Czech University of Life Sciences Prague

Faculty of Economics and Management

Systems Engineering and Informatics

Diploma Thesis

Big Data Analysis for Customer Purchase Behavior Challenges and Opportunities

Professor:

Ing. Tomáš Hlavsa, Ph.D.

Student:

Odeta Shtrepi

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Economics and Management

DIPLOMA THESIS ASSIGNMENT

Odeta Shtrepi

Informatics

Thesis title

Big Data Analysis for Customer Purchase Behaviour: Challenges and Opportunities

---

Objectives of thesis

The main objective of this thesis is to develop an in-depth analysis for identifying the best models to predict the customer purchase behaviour. While having this as a primary objective, there are also partial goals that the thesis aims to achieve such as:

- Recognizing the opportunities and statistical challenges when analysing Big Data for Customer Behaviour;
- Defining the determinant factors that influence the purchase intention and the customer behaviour in purchasing;
- Determining the possible patterns on different groups of customers based on different variables.
- Identifying the possible suitable model in predicting the scenarios of behaviour and recognizing the best one.

Methodology

To achieve the objectives, the development of this thesis involves a literature review of Big Data, of its characteristics and also the statistical challenges. These challenges with be enhanced in the practical part when analysing the dataset, each of the variables and the way the dataset is presented. Part of the literature review will also be the description of the statistical methods and techniques that will be applied on the data, such as Linear Regression, Ridge Regression and Decision Tree model. The last part of the literature consultation includes an analysis of previous papers and researches on the customer behaviour, the patterns of purchase activity and how it changes the decision making of the businesses.

The proposed extent of the thesis

60 – 80 pages

Keywords

Big Data, statistical challenges, decision making, financial risk, statistical analysis, regression analysis, predictive model

Recommended information sources

ABBOTT, D. Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst. Praha: John Wiley & Sons, Incorporated, 2014. ISBN 9781118727935.

Jianqing Fan, Fang Han, and Han Liu, Challenges of Big Data Analysis, 15 Dec 2014, Journal reference: National Science Review, 1:293-324, 2014

KOTLER, P. – ARMSTRONG, G. Principles of marketing. Harlow: Pearson, 2012. ISBN 978-0-273-75243-1.

Li, Kuan-Ching (Editor), Jiang, Hai (Editor), Yang, Laurence Tianruo (Editor), Cuzzocrea, Alfredo (Editor), Big data: algorithms, analytics, and applications, 2015, ISBN 978-1-4822-4055-9

Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed Kamaleldin Mahmoud Ali, Muhammad Alam, Muhammad Shiraz, and Abdullah Gani, Big Data: Survey, Technologies, Opportunities, and Challenges, 2014, The Scientific World Journal

NISBET, R. – MINER, G. – ELDER, J. Handbook of statistical analysis and data mining applications. Amsterdam: Amsterdam, 2009. ISBN 978-0-12-374765-5. Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money, Big Data:

Expected date of thesis defence

2018/19 SS – FEM

The Diploma Thesis Supervisor

Ing. Tomáš Hlavsa, Ph.D.

Supervising department

Department of Statistics

Electronic approval: 27. 02. 2018

prof.Ing. Libuše Svatošová, CSc.

Head of department

Electronic approval: 5. 03. 201

Ing. Martin Pelikán, Ph.D.

Dean

Prague on 29. 03. 2019

Declaration

I declare that I have worked on my diploma thesis titled "Big Data Analysis for Customer Purchase Behavior: Challenges and Opportunities" by myself and I have used only the sources mentioned at the end of the thesis. As the author of the diploma thesis, I declare that the thesis does not break copyrights of any other person.

In Prague on 29.03.2019 ______________________________

ODETA SHTREPI

Acknowledgement

I would like to thank my supervisor Ing. Tomáš Hlavsa, Ph.D. for the continuous support, advices and patience he shared with me during the working of this thesis.

Many thanks go to my friends as well who have been bearing with me during all this period and hearing all my stressful talks and complaints. Last, enormous gratitude goes to my parents, sister, and Visart who always pushed me beyond my limits and gave me unfailing support and encouragement. This accomplishment would not have been possible without their presence.

Big Data Analysis for Customer Purchase Behavior Challenges and Opportunities

Abstract

The scale of data in the past decades has shifted the way people, businesses and governments find value and make decisions. Nowadays data is the core and heart of every corporation and leveraging the analysis of it is of primary importance.

This master thesis provides a literature review of the most common and statistical challenges of Big Data as well as an overview of the most used statistical modeling and techniques for analyzing it. It also gives an overview of Customer Purchase Behavior Analysis that has been done previously. In the practical part, a dataset of Customer Purchase Behavior is used for descriptive analysis as well as predictive modeling. Some of the main predictive techniques are used to compare them with each other and to come to a conclusion on which of them is the best to use when dealing with customer purchase data.

Keywords: Big Data, Predictive Modelling, Descriptive Analytics, linear regression, ridge regression, decision tree, random forest, train set, test set.

Analýza Big Dat pro účely zákaznického nákupního chování Výzvy a možnosti

Abstrakt

Množství dat v minulých dekádách změnilo způsob, jakým lidé, podnikatelé i vlády hledají hodnoty a přijímají rozhodnutí. V současnosti jsou data jádrem všech korporací a využívání jejich analýzy je nanejvýš důležité. Tato práce předkládá přehled nejčastějších statistických výzev v oblasti Big Data i přehled nejčastěji používaných statistických modelů a postupů pro jejich analýzu. Nabízí také přehled analýz zákaznického nákupního chování. V praktické části je soubor dat týkajících se zákaznického nákupního chování použit pro deskriptivní analýzu i pro prediktivní modelování. V práci jsou porovnány některé z hlavních prediktivních postupů navzájem a zhodnocena jejich vhodnost pro použití v oblasti zákaznického nákupního chování.

Klíčová slova: Big Data, deskriptivní analýza, lineární regrese, hřebenová regrese, rozhodovací strom, náhodný les, train set, test set.

Table of content

1. Introduction 11

2. Objectives and Methodology 13

2.1 Objectives 13

2.2 Methodology 13

3. Literature Review 15

3.1 Big Data 15

3.1.1 Common Challenges of Big Data 16

3.1.2 Analytical Challenges of Big Data 19

3.2 Statistical Methods for Analyzing Big Data 22

3.2.1 Data Mining Techniques and Modelling 23

3.2.2 Choosing the right metric to enhance model performance 30

3.3 Customer Behavior Analysis 32

3.3.1 Building the complete customer purchase profile 33

3.3.2 Predicting the probability of customer purchase 36

3.3.3 Predicting personalized recommendation for each customer 38

4. Practical Part 42

4.1 Study Case Overview 42

4.1.1 Problem framing 43

4.1.2 Performance measure and assumptions 44

4.1.3 Preparing the data 44

4.2 Explanatory Data Analysis 46

4.2.1 Univariate Statistical Analysis 46

4.2.2 Bivariate statistical analysis 53

4.2.3 Correlation between the dependent variable and the numerical ones 58

4.3 Model building and training 60

4.3.1 Custom transformation of the data 60

5. Results and Discussion 65

5.1 Model building and evaluation 65

6. Conclusions 70

7. References 72

Table of content

Figure 1 The trend of data revolution [source: The Economist, based on IDC data] _________________ 16
Figure 2 Thriving in the Big Data Era [source: SAS Institute, 2012] _________________________________ 17
Figure 3 Twitter-based vs. Official Influenza Rate in the U.S. [source: You Are What You Tweet: Analyzing Twitter for Public Health. M. J. Paul and M. Dredze] ____ 18
Figure 4 Types of Analytics____ 20
Figure 5 Correlation Does Not Mean Causation [source: BigDataforDevelopment-UNGlobalPulseJune2012] ____ 22
Figure 6 CLV Regression Model [source: Five Things You Should Know About Quantile Regression] ____ 24
Figure 7 Example of kin classification [source: Medium.com] ____________________________________ 25
Figure 8 Logistic Regression Plot [source: plotly] ____ 26
Figure 9 Example of Decision Tree [source: Analytics in a Big Data World] ________________________ 27
Figure 10 Neural Network Algorithm Structure [source: Predictive Analytics for Dummies]___________ 28
Figure 11 Market Basket Analysis with Association Rules in Digital Printing [source: own] ___________ 30
Figure 12 Visual description of MAE [source: dataquest.com]___________________________________ 31
Figure 13 Confusion Matrix [source: towardsdatascience.com] __________________________________ 32
Figure 14 Outline of principles for data collection [source: Predictive Marketing]____________________ 33
Figure 15 Phases of Customer Data Collection [source: Ömer Artun Phd, Dominique Levin 2015] ______ 34
Figure 16 Data Preparation [source: Ömer Artun Phd, Dominique Levin 2015] ______________________ 35
Figure 17 Cross-Sell/Up sell Overview [source: infosysblogs.com]_______________________________ 39
Figure 18 Descriptive Analysis of the dataset [source: own] ____________________________________ 45
Figure 19 Measure of Locations [source: own] ____ 45
Figure 20 Duplicates Check [source:own] ____ 46
Figure 21 Distribution of dependent variable [source: own] ____________________________________ 47
Figure 22 Skewness and Kurtosis of the Distribution of the Dependent Variable [source: own] ________ 47
Figure 23 Numerical Variables of the dataset [source: own]_____________________________________ 48
Figure 24 Distribution of Occupation_Category [source: own] ___________________________________ 48
Figure 25 Distribution of Conjugal_Status [source: own]_______________________________________ 49
Figure 26 Product_Category_1 distribution [source: own]______________________________________ 50
Figure 27 Gender split in the dataset____ 50
Figure 28 Distribution of the variable age [source: own]____ 51
Figure 29 Dropping irrelevant columns [source: own] ____ 51
Figure 30 OptIn_Email distribution [source: own] ____ 52
Figure 31 Data Cleaning for variable Coupon_Eligible [source: own] _____________________________ 53
Figure 32 Distribution of the variable Coupon_Eligible ____ 53
Figure 33 Figure 33 Analysis of Age Category and Purchase Amount [source: own] _________________ 54
Figure 34 Total Purchase Amount per Gender [source: own]_____________________________________ 55
Figure 35 Average expenditure from each occupation category [source: own] ______________________ 55
Figure 36 Analysis between OptIn_Email and Purchase Amount_________________________________ 56
Figure 37 Coupon Eligible and Purchase Amount Analysis______________________________________ 57
Figure 38 Conjugal Status and Purchase Amount [source: own] _________________________________ 57
Figure 39 Average of Purchase Amount per Conjugal_Status [source: own] _______________________ 58
Figure 40 Correlations with the target variable [source: own] ___________________________________ 58
Figure 41 Correlation Matrix between Variables [source: own] __________________________________ 59
Figure 42 New dataset after joining train and test sets [source: own] ____________________________ 60
Figure 43 Imputing Values in Saved_Amount variable [source: own]______________________________ 61
Figure 44 Checking the data values for each variable before conversion [source: own] ______________ 61
Figure 45 The transformation of variable age into numerical [source: own] ________________________ 62
Figure 46 Transformation of variable gender [source: own] _____________________________________ 62
Figure 47 Transformation of OptIn_Email and Coupon_Eligible variables into binary ones ___________ 63
Figure 48 New train set after all the necessary transformation [source: own]_______________________ 64
Figure 49 Generic function for model building and cross-validation [source: own] __________________ 64
Figure 50 Linear Regression Performance Measure Coefficient [source: own] _____________________ 65

Figure 51 Ridge Regression Model [source: own]____ 66
Figure 52 Decision Tree Performance Measure [source: own]_____________________________________ 68
Figure 53 Random Forest Model [source: own] ____ 69

1. Introduction

Nowadays data is everything and everywhere. We find it in our daily job, no matter what is the domain we work in, in our social media, in social and health records, in Nasa's sensor's records and the list can go on for more and more lines. Statistics are stating that by 2020, the volume of data will outreach the levels of petabytes, exabytes or even zettabytes (Ismail, Information Age, 2017). In a few decades, data has gone from lacking to rich and plentiful. Almost all occupations are influenced by this massive growth of data; therefore, leveraging the analysis of it for their benefits is of primary importance.

This means that a whole new world of opportunities is offered from Big Data analytics in driving decisions of every kind and every company, in enhancing these decisions, in driving the technology revolution and in improving the overall human life. The market transformations of the last decades have empowered the embracement and usage of Big Data analytics, pushing the governments, enterprises, and institutions to reshape the way they use data.

There are also many challenges that come with the increasing of the multidimensional Big Data. Some of these challenges can be related to modeling them, to adapt and implement new architectures that address its complexity, how to capture and store it and also how to analyze and leverage the insights taken from it. A big challenge is also to equip the workforce with the right technical skills that are necessary to leverage insight from the analysis of Big Data to help business in decision-making processes.

This thesis uses Big Data as a primary source to achieve the end goal and scopes. It has the typical characteristics and issues that are arisen when dealing with Big Data. The focus of the dataset is customer purchase behavior and predicted customer behavior.

The topic of customer behavior analysis has from many years led analysists, data scientist, researchers, and decision-makers to develop hundreds and thousands of articles, papers, researches, and new methods and methodologies. In addition, all e-commerce and B2B businesses deploy prediction analysis and artificial intelligence to enhance their potential in many aspects like productivity, sales, product expansion, and many more to better fulfill the needs of the customers.

Predictive analysis of customer behavior is fundamentally predicated on being able to identify multidimensional variables that are statistically important and that give new insights on the differences between purchase behavior of the customers and their purchase intentions. There might be hundreds of variables that help come to conclusions on a prediction about customer intentions. When Machine Learning methods are applied, even thousands of variables can be included in helping the prediction to be more accurate.

The motivation for writing this thesis starts with the personal development of in-depth knowledge of the field connected to the working environment and the job role. The need to understand how Big Data is analyzed, how the insights are being produced and how data-driven should be made are at the core of this thesis.

2. Objectives and Methodology

2.1 Objectives

The main objective of this thesis is to develop an in-depth analysis for identifying the best models to predict the customer purchase behavior.

While having this as a primary objective, there are also partial goals that the thesis aims to achieve such as:

- Recognizing the opportunities and statistical challenges when analyzing Big Data for customer behavior;
- Defining the determinant factors that influence the purchase intention and the customer behavior in purchasing;
- Determining the possible patterns of different groups of customers based on different variables.
- Identifying the possible suitable model in predicting the scenarios of behavior and recognizing the best one.

2.2 Methodology

To achieve the objectives, the development of this thesis involves a literature review of Big Data, of its characteristics and also the statistical challenges. These challenges with be enhanced in the practical part when analyzing the dataset, each of the variables and the way the dataset is presented. Part of the literature review will also be the description of the statistical methods and techniques that will be applied to the data, such as Linear Regression, Ridge Regression, and Decision Tree model. The last part of the literature consultation includes an analysis of previous papers and researches on customer behavior, the patterns of purchase activity and how it changes the decision making of the businesses.

In the practical part, the analysis starts with the main characteristics and information of the specific dataset used in this thesis. After, it is necessary to apply the right cleaning and wrangling procedures so to have ‘normalized’ data before continuing with the statistical methods. During the analysis, new insights and patterns are discovered which are also part of the objectives of this work. At the end of the practical part, the three methods mentioned above are applied, and the best method is defined based on characteristics such as RMSE, ROC curve or confusion matrix. These coefficients access how all these models fit the specific problem and which one to best use in case of dealing with customer behavior type of data.

The dataset includes 500,000 records and 15 variables. It will first be split into train and test sets with a random split of 20% test. This approach is followed in all of the Machine Learning problems, and the reason behind it is to avoid the overfitting of the data. The data is overfitted if the model performs well on the train set, but it does not generalize well. (Hands-On Machine Learning with Scikit-Learn & TensorFlow, 2017) Since the main objective of this thesis is to find the best performing model in customer purchase behavior dataset, the splitting between the train and test sets helps in better analyzing the performance of the model every time new data input is present. After the data is split, the process of cleansing and wrangling is applied which leads to the univariate and bivariate distribution of the variables to find important determinants and possible patterns in the behavior, followed by the application of the methods mentioned above.

3. Literature Review

3.1 Big Data

Before we delve into more complex topics, it is of importance to understand what Big Data is. Nowadays, an immense amount of data is being generated from an immense number of different sources, most specifically from the web and electronic or smart devices surrounding us. The volume, velocity, and variety of this data make it so complicated that it surpasses the level of conventional data which is stored in typical relational databases. This has pivoted to the term of Big Data.

There are many definitions of Big Data which have also changed in different contexts and time periods. Before looking into them, anyone can ask the questions: Is Big Data defined based on the number of records or on the number of columns? Is it based on the architecture that stores it or on the non-relational databases? Does it increase with the complexity of the volume only? There are many open-questions when dealing with this topic.

The references on the definitions of such term are taken from papers of UNECE (United Nations – Economic Commission for Europe). One of these definitions' states: "Big Data is a phenomenon defined by the rapid acceleration in the expanding volume of high velocity, complex, and diverse types of data." (TechAmerica Foundation's Federal Big Data Commission). It is characterized by the exponential growth of the raw data which has overpowered the whole society. This phenomenon is both a challenge and an opportunity. It is a challenge when trying to make all the information that it offers cohesive and also building the trust in it and its source. From the other side, it is an opportunity when adding value to the businesses, governments, and individuals.

As stated by Cynthia Harvey in a post about Big Data challenges, there is no set number of gigabytes or terabytes or petabytes that separate "big data" from average-sized data (Harvey, 2017). So, what is of importance, is not primarily the number of bytes that it occupies, but all the potential for transformation that it offers, let it be in government, businesses or society itself.

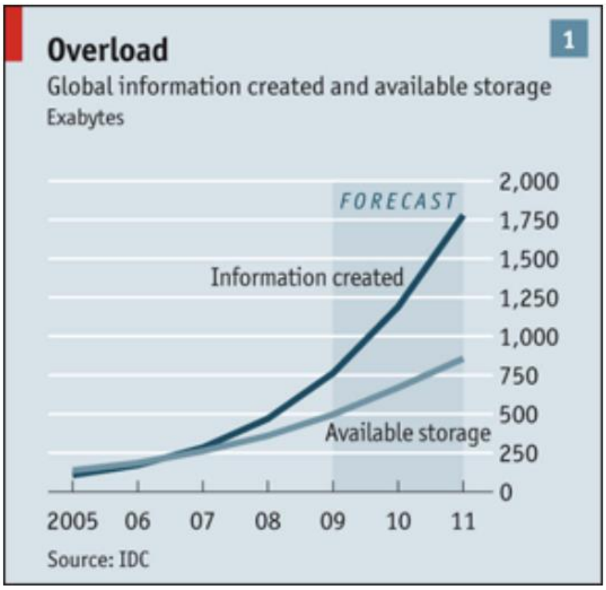


Figure 1 The trend of data revolution [source: The Economist, based on IDC data]

From the two definitions above, it can be concluded that the many definitions available change depending on the sector (private companies or governments), from the complexity, from technologies used to architecture and store it, from the size and most important from the purpose. It is evident that the interpretation over the years has changed mostly based on the volume and complexity.

Big data is everywhere, from transportation to education, from health care to fraud detection and cybersecurity, from weather to governments and social media. It surrounds us all the time, even though we might not be able to see or perceive or think about it. In our daily lives we do not deal with Big Data in most of the circumstances, and because of this, we lack the understanding of it as well as realizing the importance, challenges, and opportunities it offers. In order for people to understand the proper usage of it and to interpret it in the right way, companies and government need to make them feel safer about the usage of Big Data.

3.1.1 Common Challenges of Big Data

Before exploring the common challenges and the statistical challenges of Big Data, it is essential to understand its characteristics. They were mentioned partially in the first section, where the definitions took place. It might seem that the main characteristics of Big Data are Volume and Complexity, but in most of the studies, these characteristics are grouped in what is called "The 3 Vs. of Big Data": Volume, Velocity, and Variety. In some other studies, like the ones from SAS, they also add two other characteristics: Variability and Complexity. (SAS Insights)

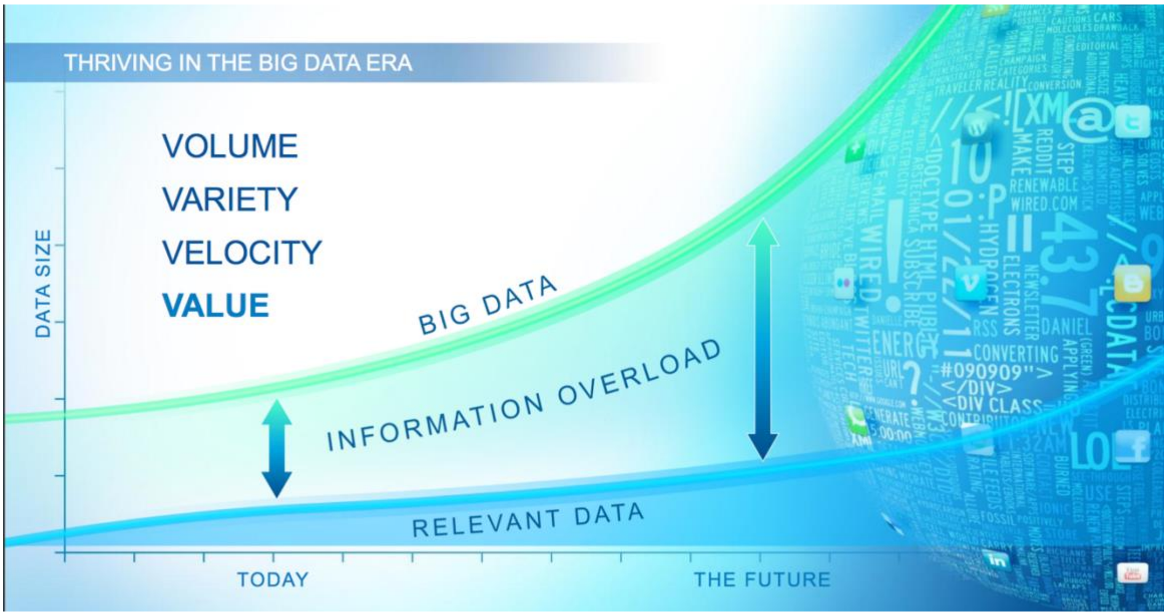


Figure 2 Thriving in the Big Data Era [source: SAS Institute, 2012]

By *Volume*, it refers to the amount and size of this data, let it be in the form of business transactions, social media information, and many more coming from a variety of sources and thanks to the technologies of these days like Hadoop, the burden of storing it, has been eased. The *Velocity* refers to the speed with which it must be dealt with. Again, thanks to the technologies of the present, real-time data is available in many sources and for different purposes. By *Variety*, it is understood that the data comes in different formats and structure, let it be denormalized and raw data, normalized and structured data and many other options. One of the most challenging characteristics when it comes to the analytics of Big Data is *Variability*. It refers to how the data flows in different periodic peaks which can be daily, monthly, quarterly, seasonal or event-trigger. The last and also most challenging characteristic is *Complexity*. It refers to the many sources combined to the different formats, records, and structures that the data is made available. This means that it has to be linked, formatted, structured and connected in such ways to make it available to use and analyze. Many common challenges are faced when dealing with Big Data which affect different areas. The first challenge has to do with *Privacy*.

Privacy can be considered as the most sensitive challenge. (UN Global Pulse, 2012) It implies social, legal and technological aspects. Privacy is first defined from the aspect of the end user or person who regulates or define the amount of information that is connected to them to be exposed. It is also defined by the companies and governments side to protect their customer and market share competitiveness and governments sovereigns. In the case of individuals, in most of the times they are unaware of the level of information they are disclosing, especially when ticking the "Consent usage" boxes without even reading how their data will be used. It is in a constant on-the-radar topic, the way giant tech companies like Facebook, Google, Twitter use individual's data, which often becomes a reason for implications and complications as well.

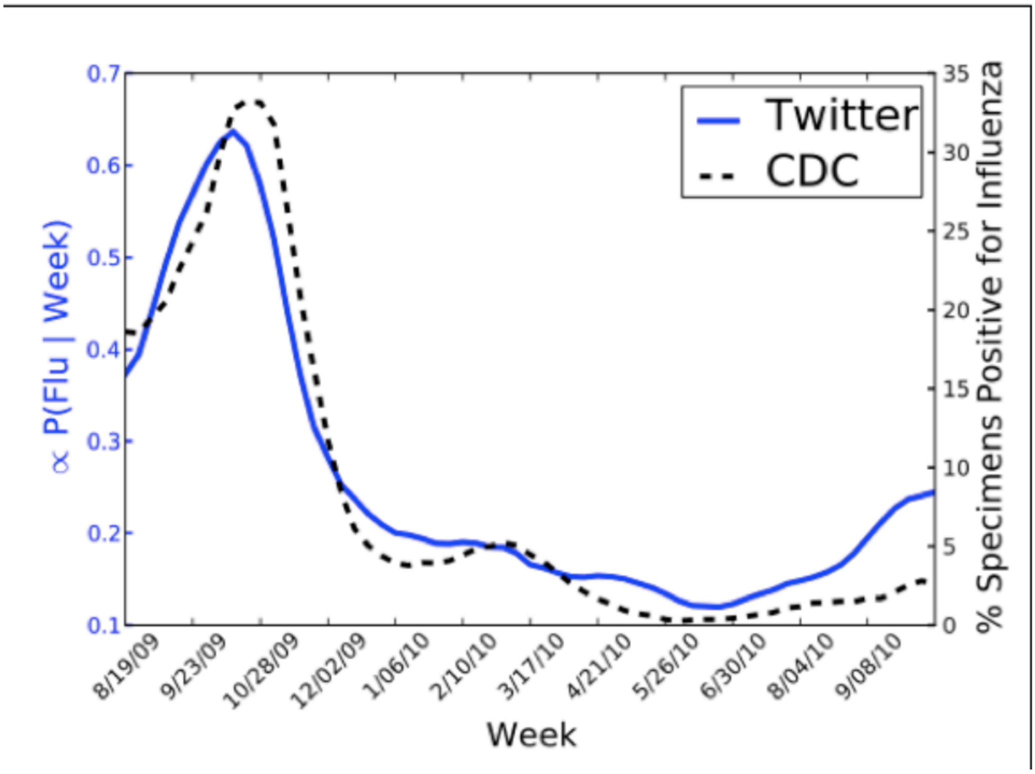


Figure 3 Twitter-based vs. Official Influenza Rate in the U.S. [source: You Are What You Tweet: Analyzing Twitter for Public Health. M. J. Paul and M. Dredze]

Another important challenge also connected to privacy is access and sharing. Even though there is a huge potential and value in the public data shared, there is far more value in the data that companies do not make available for the public. This is connected to the legal implications and people's rights considerations. Companies also tend to "keep the data secret" to avoid competitiveness. Apart all these, there are technological challenges as well which makes it often a challenge the access and the sharing of the data amongst different entities, locations, systems, etc.

There are numerous dynamic challenges especially in terms of the design and the technologies which require a deep understanding of the Big Data made available. Because new data require a deeper level of understanding, so do the designers need to capture or model the right interfaces, tooling, functionalities and the capabilities that come from that data. Part of this same challenge is also the end user who might not be that technical or familiar with the data. The designers should take into consideration the end user when designing the tooling which will make this data available. (Big Data: Issues and Challenges Moving Forward, 2013)

One of the most important dynamic challenges is the discussion Quality vs. Quantity. (Big Data: Issues and Challenges Moving Forward, 2013). There is no best decision on such question as it depends on the problem to be solved and on the domain. There are cases where having less data but with high quality is more important. These are cases where applying some predicting modeling is possible even in less quantity of data, but where all the predictors and format of the data is of good quality. Conversely, there are other situations where the quantity is significant, especially when applying Machine Learning Models. For example, if there are not enough incidents when checking the performance of a tool, system or microservice, it is impossible for a machine to learn and to predict future incidents in real time. Therefore, it depends on what the user requires and how they define which data is relevant and which not.

3.1.2 Analytical Challenges of Big Data

People who work with data in general and Big Data in particular, are familiar with the question: “What does this data tell you?” which means that the data should be available for the right statistical analysis. It is this question which drives all the decision makers within corporates or governmental institutions. Moreover, it is also this question which generates many challenges. There are also other questions that arise when analyzing Big Data, for example:

- How large can the dataset get to be able to analyze and interpret the results?

- Which portion of the whole dataset can be set as a sample and how to best define the sampling split?
- How to identify if the sample chosen is a good representative of the data?
- How to deal with missing values and which engineering changes to apply to each variable?
- How to best determine the number of indicators that should be included in the analysis?
- How to best leverage the data to answer business questions and to take actions from it?

These are general questions that whoever deals with data analytics can pose. More in details, the analytical challenges are described above.

Figure 4 Types of Analytics

Before listing and describing some of these challenges when talking about analytics of Big Data, it is also important to mention the different analytical categories. The most used ones are descriptive analytics (univariate or bivariate), predictive and prescriptive analytics. The practical part of this thesis uses a combination of descriptive analytics where all predictors are analyzed and predictive analytics as well, where different techniques and models are used not specifically to predict an outcome but to define which of these models best to use.

Moving back to the analytical challenges of Big Data, as listed in one of the papers from the 46th Hawaii International Conference on System Sciences, *Scaling* is one of the most critical issues (Big Data: Issues and Challenges Moving Forward, 2013). The more transactions are flown through the systems of companies and government, the more the data increases. By scaling, it is understood the capabilities of a specific algorithm that runs on the data to scale. As with all algorithms, after they reach their peak, they stop scaling or better say, they start decreasing. When this moment comes, it is a challenge to understand if the same algorithm can be scaled to handle more data or if it should be replaced. From the other side, scaling in denormalized and unstructured data is even more difficult.

When it comes to the analytical process itself, there are specific challenges which depend mostly on the type of analysis that needs to be conducted and also on the conclusions and decisions that need to be taken. When conducting the process of analysis, the first primary challenge is to get the right picture and the right understanding of the data. The first aspect of this is related to the authenticity and accuracy of the data. There are a few aspects related to this first challenge. For example, if a company wants to analyse the data coming from social media like posts or information relevant to the profile of the users, there is the risk that the data is false or fake. This comes as there is no guarantee that the users of the social media will produce accurate and real content for themselves and there are no strict verification steps.

Another aspect of this first challenge when analysing Big Data, is the risk of getting inaccurate perceptions and feeling regarding a phenomenon, surveys about the usage of a product, perceptions about the disease, etc. When needing to analyse such data, there will always be the risk that the data is inaccurate, and it can lead to wrong insights and conclusions. This is part of a specific type of analysis called sentiment analysis or opinion mining which aims to discover how people are happy or unhappy using a certain product, what they like or dislike, what they support or not and more. (UN Global Pulse, 2012)

A second challenge in the whole process of data analysis is the interpretation of the data gathered. Even if the data gathered is verified to be accurate, interpreting it is not always easy and straightforward. One reason can be that the sample taken may not be fully representative of the outcome that needs to be achieved. Also, it can be that the variables and indicators result weakly correlated to the sample and to be more correlated to the part of the population that is not present in the sample. The opposite case is a concern as well, where the correlation of the indicators with the dependent variable is high, but still, the sample is not accurately representative of the problem. The figure below from the UN Pulse session in 2012 shows such cases.

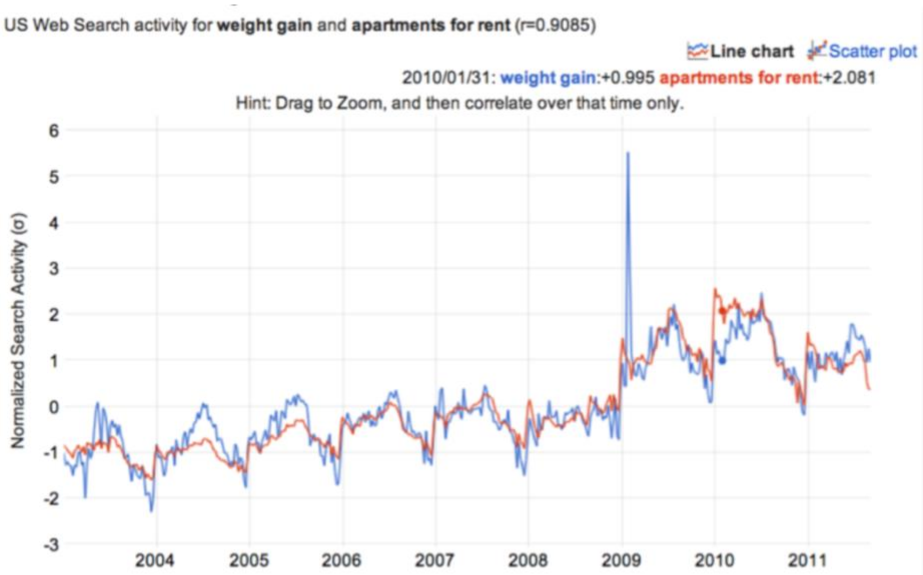


Figure 5 Correlation Does Not Mean Causation [source: BigDataforDevelopment-UNGlobalPulseJune2012]

The chart shows how highly correlated data is wrongly observed for only a sample of the dataset (people who have gained weight) instead of a more including dataset. This can lead to wrong decision making because the results are misinterpreted as being the causation of the dependent variable apartments for rent.

These two reasons demonstrate the importance of the analytical process when processing massive volume of data.

The next section describes the statistical methods and data mining techniques for analysing Big Data and specifically the most used techniques to analyse a dataset with customer transaction history.

3.2 Statistical Methods for Analyzing Big Data

In the past few decades, the increased and more complex technologies have brought the growth of data leading to higher volume, velocity, and complexity of data. This has brought immense opportunities in all fields and all industries in terms of knowledge expansion and better decision making. However, together with the opportunities, there are significant challenges, some of which have been described in the previous sections. In addition to the general challenges, there are also statistical challenges for analyzing Big Data. In order to make accurate data analysis, different categories of statistical methods have been developed. Four of them will be analyzed in detail as they are used in the practical part, but this section will give a full picture of the most used statistical techniques and modeling as well.

3.2.1 Data Mining Techniques and Modelling

Predictive modeling is one of the most used in Big Data and also differently called supervised learning. It is used in both structured and unstructured data to predict future outcomes. (Big Data for Dummies, 2013) The target variable is used for a learning process, nowadays called machine learning to predict the specific outcome. One of the most common examples where such modeling is applied is in global development for predicting life expectancy. The two most important types of predictive modeling are regression and classification. In regression, the dependent variable is continuous, whereas in classification it is categorical. Two of the most common predictive models with regression are described below: Linear Regression and Ridge Regression.

Linear regression models are the most used types to analyze a wide range of phenomena. In these types of problems, the dependent variable is usually a continuous variable with a normal distribution. (International Encyclopedia of Education (Third Edition), 2010) In these models, a relationship between the dependent variable and the independent ones (also named regressors) is built, and the linear equation built from these variables tries to fit the data. A direct example of such a model can be built between these two variables: the teen birth rate and the poverty level.

A simple linear regression model equation is built in this form:

$$CLV = \beta_0 + \beta_1 R + \beta_2 F + \beta_3 M$$

This is a typical Customer Lifetime Value problem where RFM are (recency, frequency, monetary value) (Analytics in a Big Data World, 2014) and Y is the dependent variable and X is the independent one.

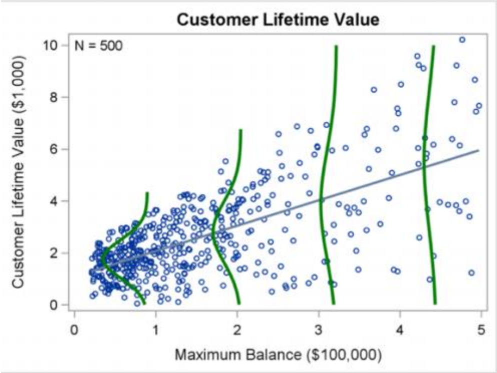


Figure 6 CLV Regression Model [source: Five Things You Should Know About Quantile Regression]

As part of the analysis, coefficients like p-value, standard errors, confidence interval are calculated. Linear regression analysis is split in Simple Linear Regression and Multiple Linear Regression. While the first one involves only one dependent variable and one regressor, the second one extends the analysis on multiple independent variables. It evaluates not only the categorical independent variables but the dimensional ones as well.

Ridge regression is just a variation of the linear regression and is used instead of it when the variables have high multicollinearity and to prevent the overfitting. In such cases, the Least Square Coefficients have an unbiased estimation with a high variance, so to reduce the standard error, the Ridge Regression model is used. There are different applications of the ridge regression, but all of them include the standardization as a first step. The standardization is applied in both dependent and independent variables by subtracting the mean and dividing by their standard deviations. This means that the ridge regression is on a standardized scale on both variables and coefficients.

Classification is the most used modeling with different types of data mining techniques. It is applied in different areas and fields from banking and marketing to education, medicine, law and many more. In banking, it can determine whether a person who applies for a mortgage score a good or bad credit risk. In education, it can determine whether a student should be placed in into regular or particular study track. In marketing, it can determine whether a customer will purchase again or will be a one-time buyer. Same as in regression, in classification, there is a target variable, but which is categorical. For example, if the target variable is the purchase amount (like in the dataset used for the practical part), it can be partitioned into predetermined classes such as low purchase, middle purchase and high purchase amount. There are other ways that the target variable can be classified, based on the effect that the most determinant regressors have in the dependent variable. The most common data mining techniques which use classification are described below.

k-Nearest Neighbor Algorithm can be used for both regression and classification problems, but it is more used for the later ones. The way it works is based on a dataset of records or data points which are divided into different classes and subsequently used to predict the classification of a new uncategorized data point. The way this is done is through feature similarity which defines how closely a data point resembles the set of features of the training set. When used for classification, the output is a class membership. The data point which has to be classified is pointed to the most common class based on the most votes from the k neighbors. The figure below is an example of the k-in classification.

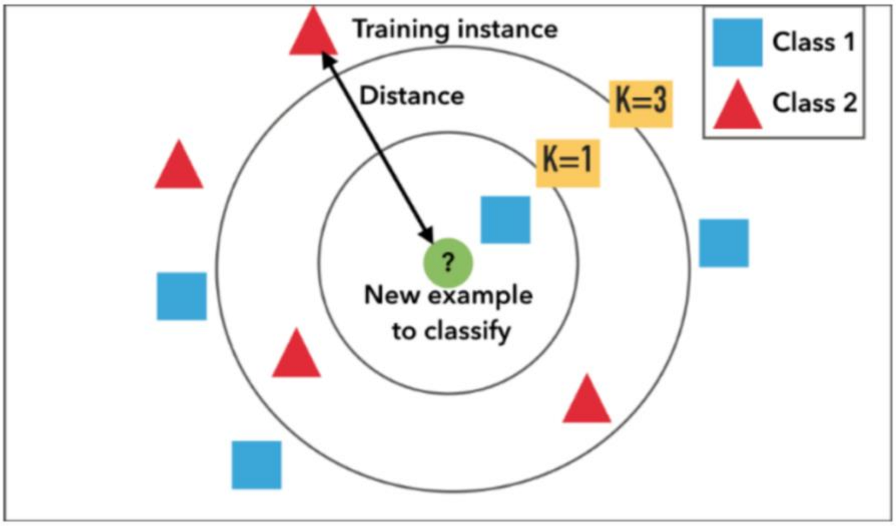


Figure 7 Example of kin classification [source: Medium.com]

Everything inside the circle is a test set, and all the data points outside are the training set. The next data point that should be classified is the green circle, and the possibilities for classification are either in the blue square of class 1 or in the red triangle of class 2. If k=3 the new data point would be classified to class 2 (as there are two triangles and only 1 square in the inner circle) and if k=5 the assigned class would be class 1(as there are three squares and only two triangles).

Logistic regression. Like all other regressions, this is also part of predictive modeling. Logistic regression very often has a discrete variable as an outcome which has two or more values. It differs from the linear regression as the target variable, in this case, is binary or dichotomous (e.g., 65-year-old and below/65 and over, absent or present, etc.). This is the reason that logistic regression is part of the classification modeling rather than regression one. It also differs in the assumptions that are placed, but then the steps applied are all the same. The equivalent of the least square method of linear regression is the maximum likelihood method in logistic regression. It returns value for the unknown parameters that maximize the probability to obtain the observed dataset. (David W. Hosmer, 2013) The mathematical form of it includes the estimation of multiple linear regression in the form of:

$$logit(p) = log\ (p_{(y\,=\,1)}\ /\ 1 - (p = 1)\ )\ ) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_n X_n$$

Some of the typical problems where logistic regression is applied are the presence (or absence) of coronary heart disease based on the age group; the prediction of low birth rate outcome based on variables such as the age of mother, smoking during pregnancy, history of premature labors, etc. Below is a simple example of how logistic regression models are usually plotted:

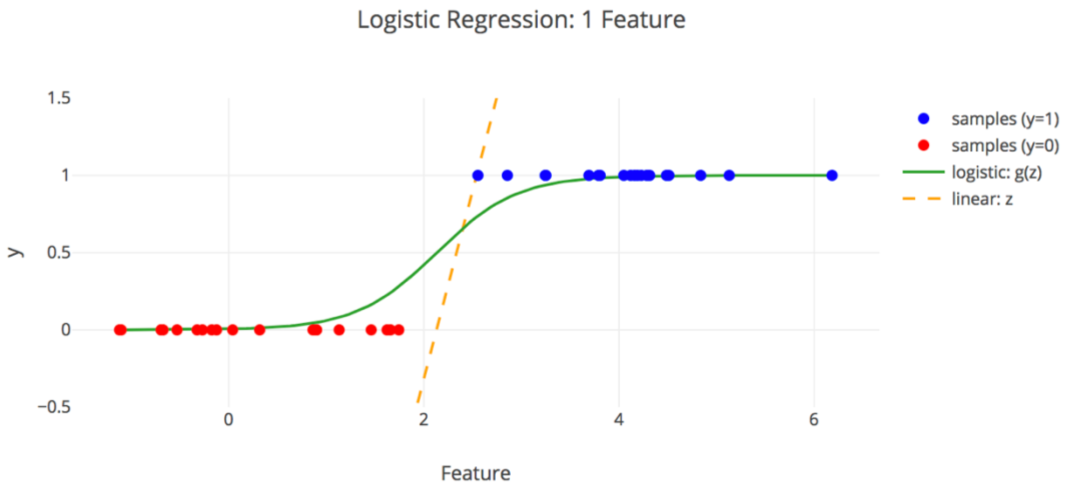


Figure 8 Logistic Regression Plot [source: plotly]

Decision Tree is a type of predictive model that is considered to be a supervised learning technique. It represents the regression or classification problem in a tree structure. As such, it always has a predefined target variable, and the input/output can both be continuous or categorical. The structure is usually a tree like where the dataset is split into two or more subsets based on the most significant variables or splitters. It includes a series of questions, the answer of which determine the split into different nodes of the tree. This model will be used in the practical part to compare it with the other methods and to decide which of them is the best one to use when predicting customer behavior.

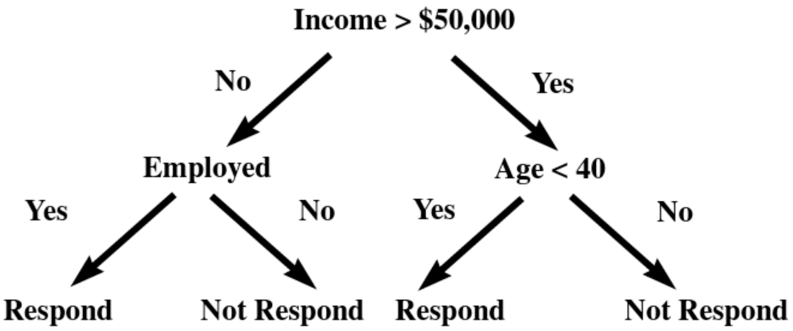


Figure 9 Example of Decision Tree [source: Analytics in a Big Data World]

When building a decision tree, there are a few decisions that need to be taken like:

- Which variable to use for splitting and at which value?
- When to stop building the tree?
- What possible values to assign to which leaf node?

Random forest is another type of supervised learning algorithm used for prediction. It is also one of the most used in Machine Learning. What is does is margining many decision trees and creating ‘a forest’ to give a better and more accurate prediction. Same as with the Decision Tree, it is also used for Classification and Regression problems which form most of Machine Learning systems. Each decision tree has access to only a random subset of training data and uses only a random subset of attributes at each node of each decision tree. Having such a wide variety of estimation from each decision tree, the random forest takes an average of all of them and makes the prediction. This is the reason that this technique performs better than one single model of the decision tree.

Neural networks are a type of predictive modeling which uses mathematical representations based on the human brain. They are sometimes considered to be a generalization of other existing statistical models. In most of the cases, neural networks are used for classification problems trying to find patterns and correlations hidden in the data and estimating future values out of them. The structure of this type of algorithm includes three layers: 1. The input layer which contains the past or present data needed to feed the next layer. 2. The hidden layer which contains functions needed to transform the input data. In the picture below, the black dots in the hidden layer represents these mathematical functions called neurons. 3. The output layer produces the predicted results based on the function on the hidden layer.

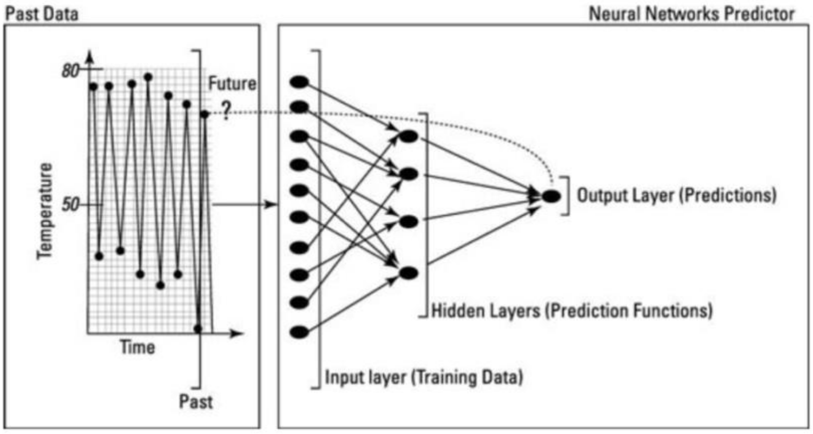

Past Data
Neural Networks Predictor
80
50
Temperature
Time
Future
?
Past
Output Layer (Predictions)
Hidden Layers (Prediction Functions)
Input layer (Training Data)

Figure 10 Neural Network Algorithm Structure [source: Predictive Analytics for Dummies]

Before moving to the descriptive modeling, there is one last technique worth mentioning as it can be used for both logistic and classification problems even though it is mostly used in the latter one. That is the *Support Vector Machine.* It is very used among a different range of problems as it has a high accuracy with low computation power needed. How it performs in simple words is based on training data which are labeled, it finds the optimal hyperplane to categorize new data points. In. 2-dimensional context, this hyperplane consists of a line which separates the plane in two parts, and therefore the new data points can be categorized in one of these classes.

Descriptive modeling is a data modeling technique which looks into the past and tries to understand what has happened. It is the most common type of analytics to describe patterns in customer behavior, and it will also be used in the practical part of the thesis. It does not necessarily have a target variable available in the dataset (i.e., customer churn or fraud detection) and for this reason, is often defined as unsupervised learning as there is no usage of the target variable to drive the supervised process. It gives a summary of all the variables present in the dataset. The most common types of descriptive modeling are association rules, sequence rules, and segmentation.

Association rules are usually used in market basket analysis or product affinity analysis. For example, if people buy business cards, there might be an interest in analyzing if these people bought envelopes as well. These types of analysis are essential to inform product placement, promotions, and marketing strategies. These rules discover the probability of co-occurrence between two or more events or items in the basket. In the figure below, the different color scale implies the different confidence by which buying a specific item implies buying the other items(s) as well. Different metrics are used to measures the strengths or quality of the association like:

- Support = Pr (A & C): the proportion of all orders that contain the antecedent(s) AND the consequent(s)
- Confidence = Pr (C | A): the probability of seeing the consequent(s) in a transaction given that it also contains the antecedent(s)
- Lift = Pr (C | A) / Pr(C): A measure of how much more often the antecedent(s) and consequent(s) occur together than would be expected if they were statistically independent. If A and C are independent, the lift score will be exactly 1.
- Leverage = Pr (A & C) - Pr(A)Pr(C): This computes the difference between the observed frequency of A and C appearing together and the frequency that would be expected if A and C were independent. A leverage value of 0 indicates independence.
- Conviction = (1 - Pr (A & C)) / (1 - Pr (C | A)). A high conviction value means that the consequent is highly dependent on the antecedent. For instance, in the case of a perfect confidence score (Pr (C | A) =1), the denominator becomes 0, and the conviction score is infinite. Similar to lift, if items are independent, the conviction is 1.

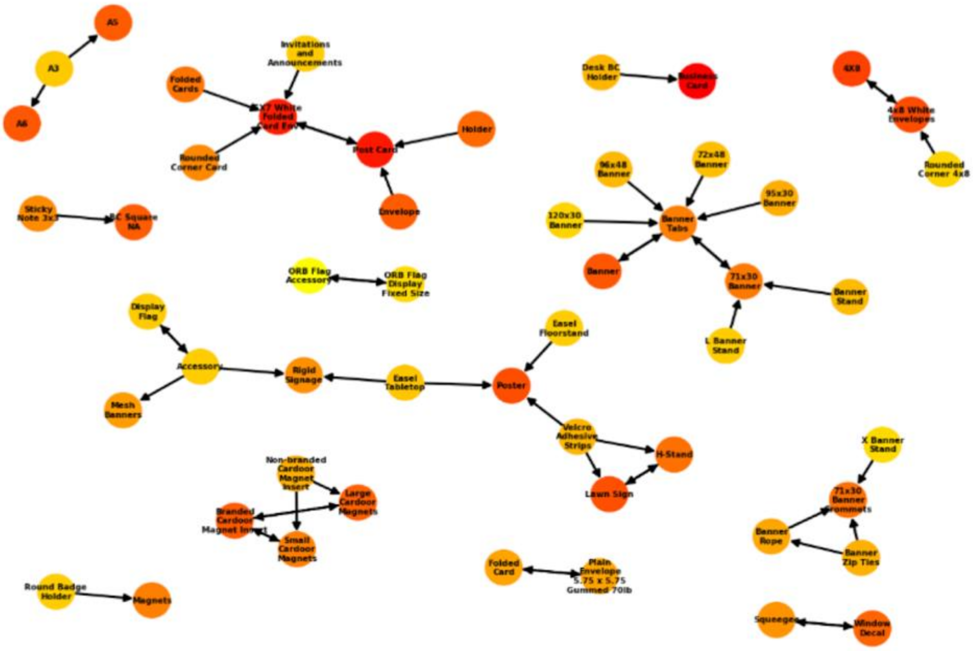


Figure 11 Market Basket Analysis with Association Rules in Digital Printing [source: own]

3.2.2 Choosing the right metric to enhance model performance

This section of the chapter is very important as based on the results of the metrics used to evaluate the model performance in the practical part; conclusions will be defined regarding which is the best model to use for the specific dataset of customer behavior. Here are some of the most used metrics to enhance the performances. The two most used metrics in the regression for continuous variables are RMSE and MAE. The one used in the practical chapter is RMSE.

RMSE (Root Mean Square Error) as the name suggests, represents the standard deviation of the differences between the predicted values (theoretical values) and the observed ones. It defines how well the model performs, and it always has a non-negative value. A value of zero would mean that the model is neither overfitting nor underfitting and that the data perfectly fits. This never happens in reality. The smaller the RMSE is, the better the model performs, but there is not a specific value range for RMSE as it depends on the depended variable. Its mathematical calculation is:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}$$

MAE (Mean Absolute Error) is the average of the absolute value of the differences between predicted values and the observed ones. It is the simplest versions of the error metric to evaluate a model.

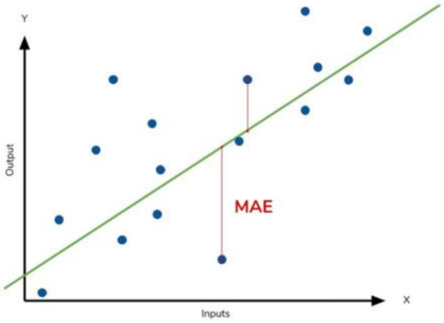


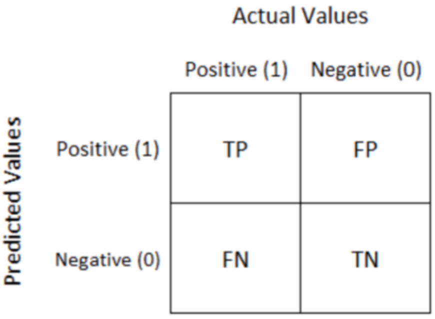
Figure 12 Visual description of MAE [source: dataquest.com]

In the graph above, the green line is a representation of the predicted model with the blue points representing the observed data points. It is very easy to understand MAE as it is just the absolute difference between the real value and the predicted one (noted with the red straight lines). Between the two metrics, even though RMSE is more complex and biased toward higher deviation, it is still the most used one among different models as it penalizes large errors more than MAE.

R^2 and Adjusted R^2 are mostly used for Explanatory data analysis to show how much the independent variables explain the target variable. It is usually better to use the Adjusted R^2 as it takes into consideration the marginal adjustment by the term added in your model.

ROC (Receiver Operating Characteristics) curve is used in predictive modeling to distinguish the difference between correct positive and false favorable rates. So, to do this, the model has to include both true positives and true negatives. The way the ROC curve does this is by plotting the sensitivity in the y-axis; the probability of predicting that a true positive is a real true positive and by plotting the specificity in the x-axis; the probability that a true negative is as such.

Confusion Matrix is a performance measure for classification problems where the output can be 2 or more classes. The four possible results are a combination of predicted and actual values:



Predicted Values	Actual Values: Positive (1)	Actual Values: Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Figure 13 Confusion Matrix [source: towardsdatascience.com]

The actual values are named as True/False and the predicted values are named as Positive/Negative.

3.3 Customer Behavior Analysis

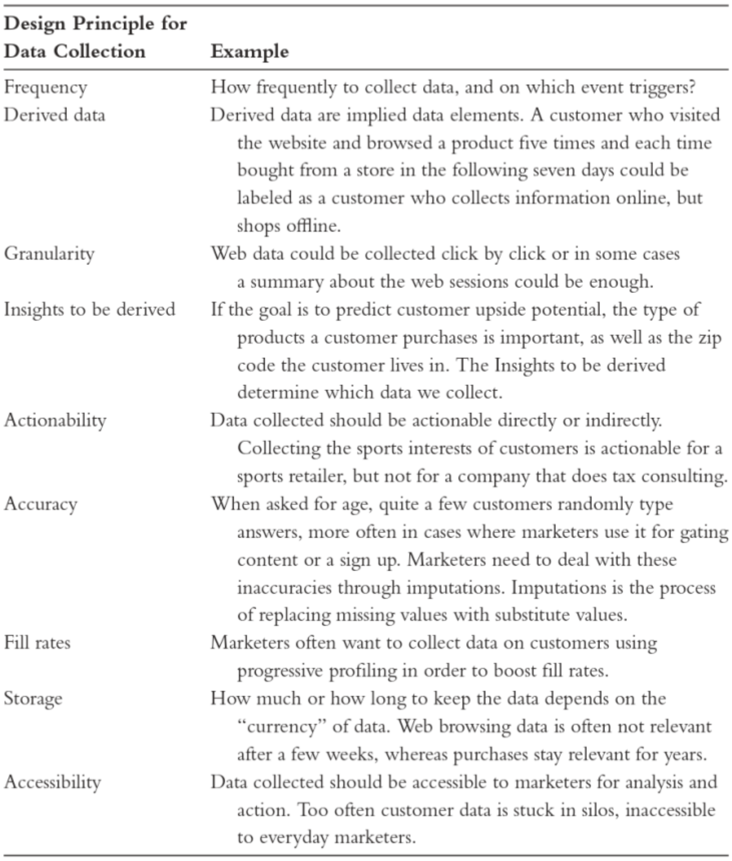
Even though there are many technologies and approaches put in place for customer behavior analysis, it is still considered to be not a fully explored market with more potential to come. Big Data in such area gives to the businesses new insights by analyzing and transforming it. (Khade, 2016) By analyzing the customer behavior, the companies can improve their sales decisions, enhance and improve marketing, depict fraud and detection and many more applications. There are many predictive approaches in analyzing customer behavior, and the most used one is the Decision Tree. In this thesis, also other techniques will be used like Linear Regression, Ridge Regression, and Random Forest.

In the last couple of years, customer behavior analysis and prediction are being labeled as predictive marketing as all the insights are valuable mainly and initially for marketing decision from a data-driven perspective. The rise of such term is encouraged from a few factors as more and more technologies are available to fulfill such analytics and to give the marketers and salespeople the right answers; the customers are demanding more and more personalized content and personalized offers and wider range of products as consumerism has reached its peak worldwide; those who have already adopted such predictive analytics can show the added value to their businesses. (Ömer Artun Phd, 2015)

3.3.1 Building the complete customer purchase profile

Before starting with what predictive modeling for customer behavior includes, it is important to get a complete picture of the customer purchase behavior. This includes different steps of aggregation, cleaning, and wrangling and then analyzing it.

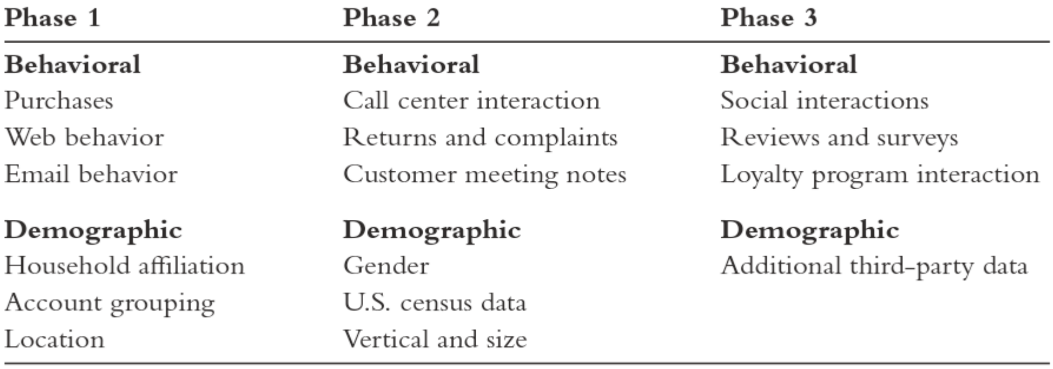
Data collection: It all starts with understanding how much data is needed to collect. As suggested in the book from (Ömer Artun Phd, 2015), the best data comes directly collected from the company, called first-party data. Data collected outside of the company is called third-party data and many times has privacy issues with breaching the boundaries. Also, there are additional costs and challenging when collecting third-party data. Nowadays, first-party data includes more historical and real-time data than ever before. The figure below shows an outline of the design of the data that needs to be included when completing the customer profile.



Design Principle for Data Collection	Example
Frequency	How frequently to collect data, and on which event triggers?
Derived data	Derived data are implied data elements. A customer who visited the website and browsed a product five times and each time bought from a store in the following seven days could be labeled as a customer who collects information online, but shops offline.
Granularity	Web data could be collected click by click or in some cases a summary about the web sessions could be enough.
Insights to be derived	If the goal is to predict customer upside potential, the type of products a customer purchases is important, as well as the zip code the customer lives in. The Insights to be derived determine which data we collect.
Actionability	Data collected should be actionable directly or indirectly. Collecting the sports interests of customers is actionable for a sports retailer, but not for a company that does tax consulting.
Accuracy	When asked for age, quite a few customers randomly type answers, more often in cases where marketers use it for gating content or a sign up. Marketers need to deal with these inaccuracies through imputations. Imputations is the process of replacing missing values with substitute values.
Fill rates	Marketers often want to collect data on customers using progressive profiling in order to boost fill rates.
Storage	How much or how long to keep the data depends on the "currency" of data. Web browsing data is often not relevant after a few weeks, whereas purchases stay relevant for years.
Accessibility	Data collected should be accessible to marketers for analysis and action. Too often customer data is stuck in silos, inaccessible to everyday marketers.

Figure 14 Outline of principles for data collection [source: Predictive Marketing]

The first thing to keep in mind when collecting data is *the end goal.* If the company needs to predict if a specific product extension will be successful, what first needs to be collected is historical data of purchases of the previous version of the product. Other important things to keep in mind are the amount of data to collect and the type of it. Many marketers make the mistake of collecting an enormous amount of data and a huge variety of attributes and variables. Sometimes it is enough to start with just a small volume of data and to select the most important regressors. In terms of what data to collect, historically all marketers have collected purchase data and demographic data of the customers. Nowadays more complex data is being collected like behavioral data which results in more temporal information. These data include what the customer clicked, what product they viewed, bid, bought, returned, how many times was a specific product bought and for which product they wrote reviews and many more. The figure below shows how to divide a phased data integration and what type of data to include. This structure is built with the end goal to increase customer engagement and customer lifetime value. The most essential behavioral data include purchase (purchase amount or purchase value), web behavior and email behavior. Some of the most important demographic data include age, gender, marital status, the domain of the company, how many years the customer has lived in that city (if the analyze is focused on the in-store customer purchase behavior), and many more. Even if there is only behavioral data included, it already gives enough information. For example, the purchase generates much meta-data like the amount of purchase and value, item information and transaction information.



Phase 1	Phase 2	Phase 3
Behavioral	**Behavioral**	**Behavioral**
Purchases	Call center interaction	Social interactions
Web behavior	Returns and complaints	Reviews and surveys
Email behavior	Customer meeting notes	Loyalty program interaction
Demographic	**Demographic**	**Demographic**
Household affiliation	Gender	Additional third-party data
Account grouping	U.S. census data	
Location	Vertical and size	

Figure 15 Phases of Customer Data Collection [source: Ömer Artun Phd, Dominique Levin 2015]

Data preparation: When asked Data Scientist on what the most important thing is when dealing with data analysis, they will answer data wrangling. Moreover, indeed, data cleaning and preparation takes up to 95% of all work done. If the customer profile is not complete and data is missing, the results may be misleading.

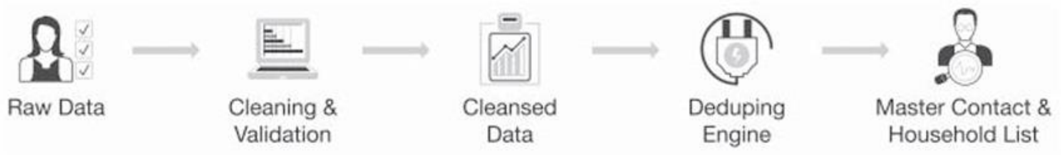


Figure 16 Data Preparation [source: Ömer Artun Phd, Dominique Levin 2015]

The figure above gives a high-level overview of the steps in data preparation. Each of them will be discussed in detail. After receiving the raw data, each of the variables needs to be validated through validation checks. Dimensions such as email address, billing address, phone number, and all other demographical information. Many software can normalize missing or misleading data such as with tagging or replacing abbreviations in customer names. Other types of validations for email addresses include syntax validation or invalid email filtering.

After cleaning the missing and incorrect data and after validating it, the deduplication comes in place. This is a technique used to eliminate duplication of records and it is essential as it avoids targeting the same customer more than once and also to increase the accuracy of the metrics and critical results. Such a process can be done through fuzzy algorithms which based on some similarity logic and threshold, defines if two records belong to the same customer or not.

After collecting and cleaning the data, the information is ready to be used for customer analytics and to create powerful meaning from it. In most of the cases, the insights start being evident after the data has been cleaned and formatted in such a way that you can start answering questions. For example, when all of the demographic information is available and cleaned, questions such as: which gender has the biggest share of a specific market, which age category spends more, are single or married people more willing to buy, etc., can lead to possible answers. So, with all the possible demographic data available and cleaned, there is the possibility to uncover the target variable that is being analyzed and helping marketers change their marketing strategies. The next sections are going to cover predictive modeling in different aspects of the customer behavior and purchase journey, whether a likelihood for buying or predicting individual recommendations or predictive modeling to grow customer value. The theoretical part covers only one of this specific behavior, which is the prediction of the purchase amount that each customer is willing to spend against different product categories. This will help marketers create more personalized content.

3.3.2 Predicting the probability of customer purchase

With probability or likelihood to buy, a whole range of behaviors is included. This can imply whether a customer that is searching or marking it as favorite or bidding for a specific product is going to buy it. These studies help marketing be more focused on data-driven decisions leading to specialized discount per each customer and therefore increasing the sales and retention rates. In the dataset used in the practical part, there is information included about specific discounts for a specific user and how much money the customer would save on a given product if they would use the coupon given. The first step in a probability to buy prediction model, is to analyze customers who are not customers yet and who have not made their first transaction. They would be analyzed based on their searching history in the e-commerce of the company, on how they interact with the promotional emails they receive, whether or not they click on the links that the emails include, etc.

The second step and a more natural modeling is the prediction to buy for already exciting customers. The reason this is easier is because there is already enough information for people who have already bought at least once. The model uses previous purchases and transactions of the existing buyers and derives mathematical constructs from building the modeling. There are a few approaches that are helpful to analyze the customer behavior of exciting buyers:

- *The RFM approach* which stands for recency, frequency, and monetary value. This technique analyzes the buyers who have bought more recently from the company or who have bought more frequently or who have spent the most money at that company. All these types of customers are more likely to spend again than other buyers. (Optimove, 2018) RFM has the advantage that it is easy to implement and understand also from the marketers and managers and does not require the implementation of any tool. What it lacks is the high level of accuracy in predicting future behavior as it only describes the past events of the buyers. It also has another disadvantage that it takes into consideration only a specific point in the past and not a whole past customer behavior.

- *Customer segmentation* which splits the customers of a company into different subsets or different segments based on a specific target behavior that is common for all the buyers of that segment. This avoids the risk of losing important behavior for each specific customer which comes with the aggregation of the data that is done in the RFM approach. Some of the other indicators that help marketers define the segmentation with higher accuracy are the demographic and psychographic indicators which together with the behavioral ones, define the complete profile of each buyer.
- A deeper approach to the standard segmentation is the *micro-segmentation* which are a very small group of buyers created from different factors and indicators, including even behavioral predictions. Each of these micro-segments is extremely precise in terms of what characteristics it includes. Micro-segmentation is extremely important for digital companies as the personal experience each buyer has with the offerings of the company affects immediately how they are segmented and therefore the percentage of outselling from these personalized contents. As Gartner predicts, the companies who excel in segmentation and personalization, outsell companies that do not by 20%. (Unruh, 2018)
- *Lifecycle stages and segmentation layers* is a combination between a higher level of lifecycle stages and deeper(micro) segmentation. An example of this approach would be the customers of an online gaming company. Initially, they can be segmented into different lifecycle stages such as new, expert, churn and reengaged and then a deeper micro-segmentation may be applied. In most of the cases, this deeper segmentation layer includes cluster analysis based on indicators that share common patterns or context. As in the generic segmentation, the cluster analysis can be applied based on behavioral attributes or demographic ones.

All of the above techniques use past historical data of customers for explanatory data analysis and possible prediction. Even though the practical part will focus only on this specific type of prediction, so the probability of customer purchase, there is also another type of prediction analysis and modeling that can be done on a customer profile such as:

predicting personalized recommendation for each customer, predicting customer value and customer journey etc.

3.3.3 Predicting personalized recommendation for each customer

Sending personalized content to customers comes together by analyzing their behavior and demographic information. After a customer has been offered a personalized product, the recommendation based on the reaction for that product is a tactic that nowadays most companies follow. The three steps of this predictive modeling are: sending the right personalized recommendation at the right time, understanding the context and sending the content. (Ömer Artun Phd, 2015) The first is exactly sending the right recommendation at the right time. Moreover, the right time can be anytime from the moment the customer is buying, to after buying, or during his lifecycle or when the company has not heard back from them in a while. These are referred respectively as upsell, cross-sell and next sell. An example of an upsell when about to finalize a purchase in Amazon website, and a recommendation next to your basket pops out suggesting you buy a bigger version or a complete version (therefore more expensive) of that product. The upsell recommendation is mostly tied to the product rather than the customer. Same as the upsell recommendation, even cross-sell is done during the purchase process and is tied up to the product cluster. It recommends you buy a product which shares some affinity score with the product currently in the basket. Not necessarily these two techniques are tied to the product only. For example, when the ticket price comparison engine Kayak, launched a pricing prediction tooling telling the customer whether to make the purchase or to wait for possible dropouts on the price, this brought a great customer experience.

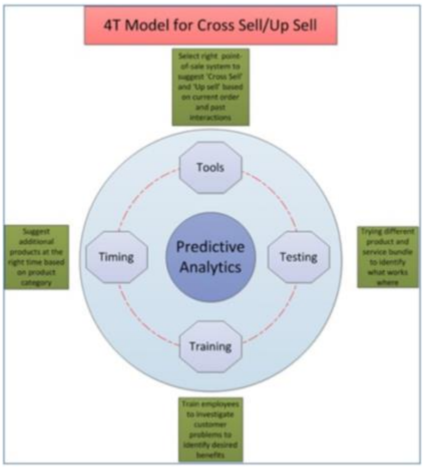

4T Model for Cross Sell/Up Sell
Select right point-of-sale system to suggest 'Cross Sell' and 'Up sell' based on current order and past interactions
Tools
Suggest additional products at the right time based on product category
Timing
Predictive Analytics
Testing
Trying different product and service bundle to identify what works where
Training
Train employees to investigate customer problems to identify desired benefits

Figure 17 Cross-Sell/Up sell Overview [source: infosysblogs.com]

From the other side, there is also a recommendation modeling technique made after the purchase called next-sell. This can be more accurate as it takes into consideration not only the specific product bought at the moment of the last transaction but also the whole possible transaction history of that customer and their demographical data. So, this technique is not based mainly on the product, but in the customer.

The recommendation made during the customer lifecycle is essential and must be modelled properly as they can make a massive impact on re-engaging lapsed customers. Before sending these recommendations, it is necessary that some initial predictive analysis has been applied to the customer lifecycle so to predict the buyers at risk of leaving. After identification, the personalized recommendation can be sent either via email or directly to the website as a message or as a coupon/discount. With all the automation marketing tooling that have been developed in the last years, it is very straightforward to use the same HTML code, or email content for all lapsed customers with the dynamic fields of product/buyer name and all other contents that need to be personalized. In the dataset used from the eCommerce company, there are discounts offered to different customers and a specific product recommendation so that the customer can use that coupon. The way this selection is done is through customer identification algorithms which take into consideration the previous purchases of them and the product affinity with these purchases.

In predicting model for the personalized recommendation, it is very important to understand the customer context. For example, if a customer who usually watches horror movies, is trying to find a proper movie to watch with their kid, the right recommendation should be sent to them, which is relevant to that specific context.

When it comes to context, it is definitely applicable to the product side as well. If the recommendation engine is trying to find an additional accessory for an electronic device, it should check which was the last electronic product the customer bought and what accessory might fit that product. This is very important also for the reason that it avoids cases when the product recommended is not something that you might have already purchased. Also, it avoids cases when the product offered is entirely out of context with your personas: if your shoe size is 35 and the personalized email says something like: “The right shoe with the right size for you”, but once you open the email and click on the link, you find out that the shoe size offered is 40, then this is entirely out of sync and not applicable at all. However, these cases do happen. Also, if the e-commerce is sending a personalized content, then it has to be a real one and not implied only on the title of the email, but then the link redirects to the generic website of the company. These cases also happen as well. All of the cases mentioned in this section have happened, and they are not good customer experience. Recommendations in the right context are the key success factor of business in general and e-commerce in particular.

After paying the right attention to the context, the final step is to pay the right attention to the content. It has been proven from various companies that the click-through rate of the basic emails with personalized content is three times higher than the click-through rate of the more beautifully designed emails, but with less meaningful content. (Ömer Artun Phd, 2015)

In the context of choosing the right subject or theme, it is very important to give customers a certain level of control over the product or service that is being is recommended. The news that the company Target sent personalized emails to pregnant women before they had even shared the news with their loved ones made a big fuzz about the privacy and controll the customers have over their transaction history and is used as an example when comparing good and bad predictive analyzing models. (Hill, 2012) From the other side, even the marketer themselves need a high amount of control over the products that they are recommending. For example, they would not want to send recommendations or offers about a product that is out of stock or that are proven to have low quality or that might be expired.

All of the merchandising rules has to be merged with the algorithms applied for predictive analyzing.

These and many more approaches, techniques and best recommendations are available nowadays when implementing marketing strategies and decisions. What is of significant importance and similar for all companies are the topics covered in this section, so: initially to build the full profile of the customer including behavioral, psychological and demographic information. The next step is to start building the predicting modeling based on the available dataset. The techniques can vary from descriptive analytics only to analyze past historical data to predictive analytics including regression or classification data mining techniques to be more proactive. One of the ways that the analytics team or data scientists in collaboration with marketers can decide for the best technique to implement, is by comparing the right metrics in terms of model performance. This is one of the sections included in the practical part. Once the decision has been made, the customer purchase can be analyzed and predicted with the right model and technique. After such an important process, the prediction might include additional steps such as determining the right recommendation to be sent to the right segment of customers or the right buyer. For this, it is important to understand when the right time for such recommendation is, what is the context of the product or buyer themselves and what specific content to send to specific users.

4. Practical Part

4.1Study Case Overview

The dataset has been taken by direct-observation from an e-commerce company. After having granted legal right to use this dataset for this research thesis, the only condition was to give to the company an anonymous identity, so it was renamed as “ABCommerce”. It was clearly stated since the beginning that the motivation behind requesting this dataset was for academic purposes only and that this information will not be used for any other purposes. It includes 500.000 records with various information about customer profile (both demographic and behavioral information). Before analyzing the data and its variables, it is important to understand the benefit of this study. Imagine that ABCommerce Marketing department needs to lunch new and personalized marketing campaigns based on the results of this analysis. Therefore, the offers have to vary. In order to do this, the company needs to create a full customer profile for all of the customers and to analyze it using descriptive and predictive analytics. After having this analysis in place, it is then possible to predict furthermore for what kind of recommendation or personalized marketing campaign to send to each of the buyers.

During the step by step analysis, it will be possible to identify some of the challenges of Big Data that were mentioned in the literature review. Also, different actions will be taken to solve some of these issues. After discovering these challenges, different statistical methods will be applied to the data set for the primary goal, that is to come to a conclusion on which of the techniques is the best one in predicting the number of purchases that the customers will make on different product categories. That means that this is a prediction model on how much the customer will spend. Also, secondary insights will be gained during the whole process. To achieve the main goal, Machine Learning model will be used. Machine Learning is the science (and art) of programming computers so they can *learn from data*. Why machine learning for this specific dataset and goal? Applying ML techniques to dig into large amounts of data can help discover patterns that were not immediately apparent. This is called *data mining*. (Géron, 2017). The different models used, will be able to learn from this data set and from there, using performance measure metrics, a conclusion can be made on which of the models is the best one when dealing with customer purchase datasets.

The dataset is cross-sectional as the observations for the customers are gathered at the same point in time and consists of comparing the differences between the subjects.

There are 14 variables from which 11 are qualitative, and only 3 are quantitative (Saved_Amt, Purchase_Amount and Stay_In_Current_City_Year). These variables will be analyzed in detail in the further sections. The technology used for this analysis is Jupyter Notebook with Python. Jupyter is an open-source software where Python code can run and make the results visible within the notebook.

4.1.1 Problem framing

Before looking at each variable and how they affect the depended one, it is important to understand how ABCommerce will take advantage of this analysis. Therefore, it is important to define the type of system, the algorithms that will be applied to the dataset, the right performance measure and the assumptions that need to be verified.

Machine Learning systems can be classified according to the amount and type of supervision they get during training. There are four major categories: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. (Géron, 2017). Because the training set of this dataset includes the solutions which are called label, this system is a supervised system. Moreover, because the dependent variable to be predicted is a numeric variable (the amount of purchase the customer will spend at a specific time on the website), this task is a regression task. It can also be changed to a categorical task, which consists of changing the dependent variable into a categorical one to define whether the customer will purchase in the future from that specific product category or not.

The algorithms used for this study are Linear Regression, Ridge Regression, Decision Tree and Random Forest which have been described theoretically in the literature review.

There is also another categorization of the systems based on the learning ability (if the system can learn incrementally from a stream of incoming data or not). The specific system of this study does not have a continuous stream of data; therefore, it is considered to be a batch learning system.

4.1.2 Performance measure and assumptions

Since this is a typical regression model, the most important metric of such model is the Root Mean Square Error (RMSE). RMSE, as mentioned in the theoretical part as well, is the standard deviation of the residuals and it measures the level of error that the system will make while predicting the dependent variable.

One of the main challenges in Big Data analysis is formulating the assumptions and verifying them. From the other side, this also can help to catch issues in the early stages. For this specific system, some of the assumptions that can be made are:

- Which age category spends more and is more willing to purchase?
- Which gender spends more and is more willing to purchase?
- Do customers who are opted-in for the email marketing communications have a higher tendency to spend?
- Do buyers who are eligible to a coupon have a higher tendency to spend more?
- Is the amount of years lived in a city relevant variable for a dataset of purchases made online?
- Single people tend to spend more than married people.
- Customers from different occupation categories have different tendencies to spend on different product categories.

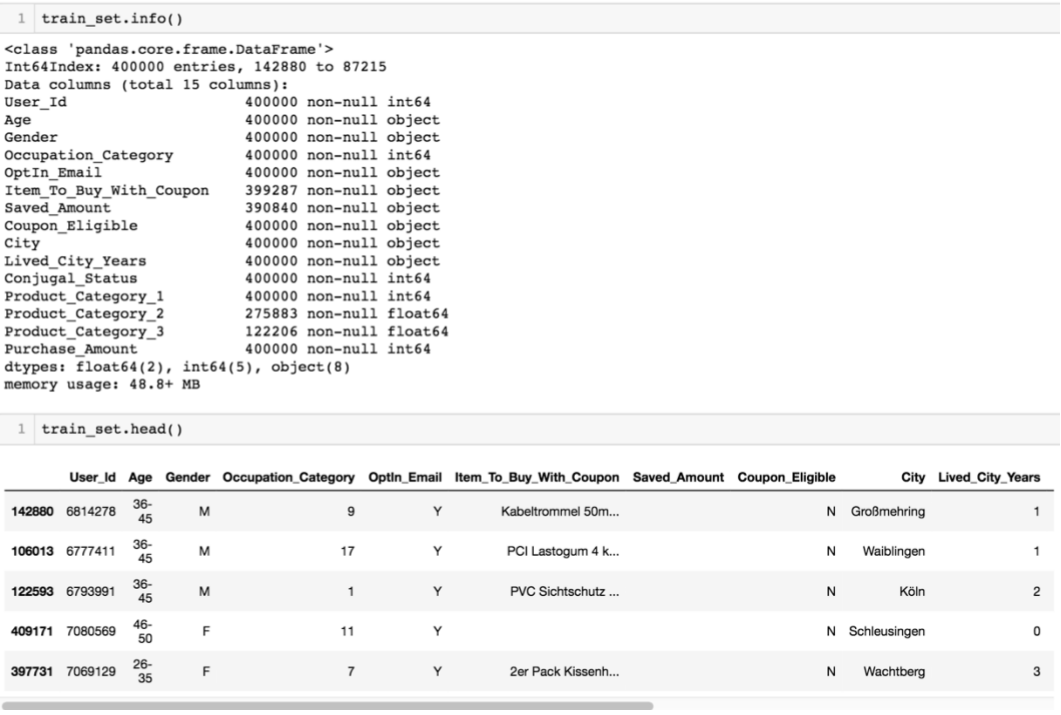
There are also other possible assumptions that can be made based on the available dataset.

4.1.3 Preparing the data

One of the main challenges before preparing a dataset for machine learning algorithms is to split the dataset between train and test set. The simplest way is to randomly pick 20% of the dataset and set it aside. The disadvantage of this is that whenever the program runs, it takes different test sets, until it uses the whole dataset. However, for simplicity and study purpose, the test set here is defined by randomly picking 20%.

This is done by creating a train/test split function and applying it to the dataset. Since this has the drawback that generates new test and train set every time it runs, the data frame train and test will be saved as CSV files and used in the next steps of the analysis. From this step, all the analysis will be done on the train_set data frame.

The train set contains 400,000 records and 15 variables. Purchase Amount is the dependent variable, and all other 14 variables are independent.



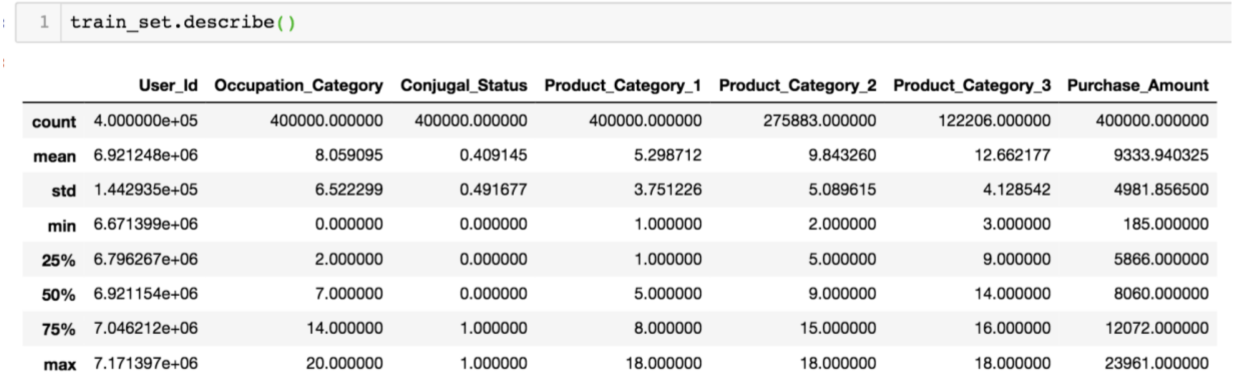
```
train_set.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 400000 entries, 142880 to 87215
Data columns (total 15 columns):
User_Id                    400000 non-null int64
Age                        400000 non-null object
Gender                     400000 non-null object
Occupation_Category        400000 non-null int64
OptIn_Email                400000 non-null object
Item_To_Buy_With_Coupon    399287 non-null object
Saved_Amount               390840 non-null object
Coupon_Eligible            400000 non-null object
City                       400000 non-null object
Lived_City_Years           400000 non-null object
Conjugal_Status            400000 non-null int64
Product_Category_1         400000 non-null int64
Product_Category_2         275883 non-null float64
Product_Category_3         122206 non-null float64
Purchase_Amount            400000 non-null int64
dtypes: float64(2), int64(5), object(8)
memory usage: 48.8+ MB
```

```
train_set.head()
```

	User_Id	Age	Gender	Occupation_Category	OptIn_Email	Item_To_Buy_With_Coupon	Saved_Amount	Coupon_Eligible	City	Lived_City_Years
142880	6814278	36-45	M	9	Y	Kabeltrommel 50m...		N	Großmehring	1
106013	6777411	36-45	M	17	Y	PCI Lastogum 4 k...		N	Waiblingen	1
122593	6793991	36-45	M	1	Y	PVC Sichtschutz ...		N	Köln	2
409171	7080569	46-50	F	11	Y			N	Schleusingen	0
397731	7069129	26-35	F	7	Y	2er Pack Kissenh...		N	Wachtberg	3

Figure 18 Descriptive Analysis of the dataset [source: own]

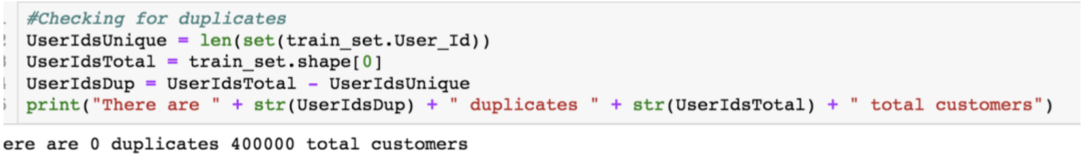


```
train_set.describe()
```

	User_Id	Occupation_Category	Conjugal_Status	Product_Category_1	Product_Category_2	Product_Category_3	Purchase_Amount
count	4.000000e+05	400000.000000	400000.000000	400000.000000	275883.000000	122206.000000	400000.000000
mean	6.921248e+06	8.059095	0.409145	5.298712	9.843260	12.662177	9333.940325
std	1.442935e+05	6.522299	0.491677	3.751226	5.089615	4.128542	4981.856500
min	6.671399e+06	0.000000	0.000000	1.000000	2.000000	3.000000	185.000000
25%	6.796267e+06	2.000000	0.000000	1.000000	5.000000	9.000000	5866.000000
50%	6.921154e+06	7.000000	0.000000	5.000000	9.000000	14.000000	8060.000000
75%	7.046212e+06	14.000000	1.000000	8.000000	15.000000	16.000000	12072.000000
max	7.171397e+06	20.000000	1.000000	18.000000	18.000000	18.000000	23961.000000

Figure 19 Measure of Locations [source: own]

Before starting the EDA, it is important to check the duplicates. In case there are, they need to be removed by creating a number of observations per customer.



```
#Checking for duplicates
UserIdsUnique = len(set(train_set.User_Id))
UserIdsTotal = train_set.shape[0]
UserIdsDup = UserIdsTotal - UserIdsUnique
print("There are " + str(UserIdsDup) + " duplicates " + str(UserIdsTotal) + " total customers")

ere are 0 duplicates 400000 total customers
```

Figure 20 Duplicates Check [source: own]

From the figure above, there are no duplicates in the train set and even in the whole dataset as the selection was made on purpose to pull only distinct user_ids.

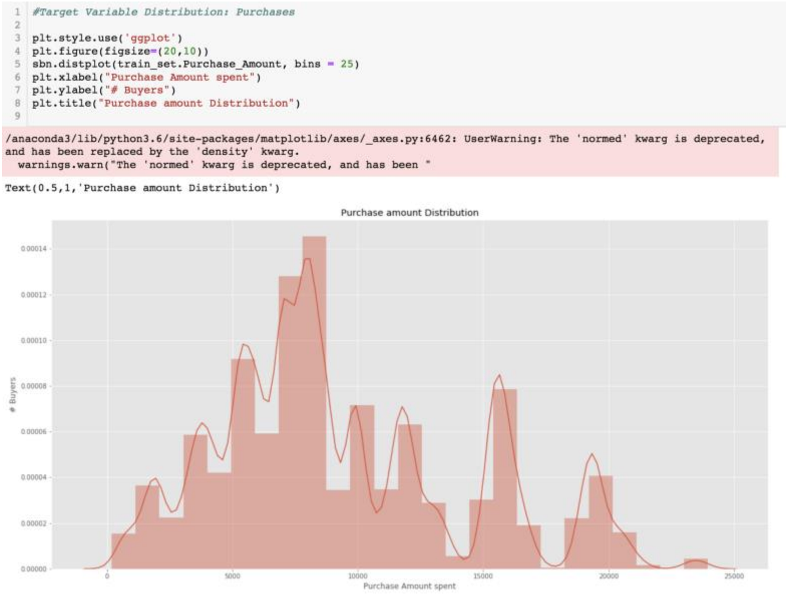
It is also important to define the type of variables that are present so to understand their influence on the dependent one and also to determine the statistical methods that will be used. The categorical variables of the dataset are: User_Id, Age, Gender, Occupation_Category, OptIn_Email, Item_To_Buy_with_Coupon, Coupon_Eligible, City, Conjugal_status, Product_Category_1, Product_Category_2, Product_Category_3. The quantitative variables are Saved_Amount, Lived_City_Years, Purchase_Amount.

4.2 Explanatory Data Analysis

4.2.1 Univariate Statistical Analysis

Before preparing the data for pre-processing, it is essential, to get a full understanding of the variables and how each of them affects the dependent one. This is done through the EDA techniques. Initially, each variable will be analyzed independently through the Univariate Statistical Analysis. (The SAGE Encyclopedia of Communication Research Methods, 2017)

Dependent variable: Purchase Amount



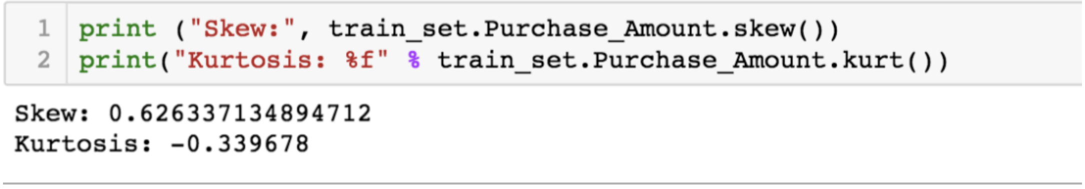
```
#Target Variable Distribution: Purchases

plt.style.use('ggplot')
plt.figure(figsize=(20,10))
sbn.distplot(train_set.Purchase_Amount, bins = 25)
plt.xlabel("Purchase Amount spent")
plt.ylabel("# Buyers")
plt.title("Purchase amount Distribution")

```

```
/anaconda3/lib/python3.6/site-packages/matplotlib/axes/_axes.py:6462: UserWarning: The 'normed' kwarg is deprecated,
and has been replaced by the 'density' kwarg.
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```

```
Text(0.5,1,'Purchase amount Distribution')
```

Figure 21 Distribution of dependent variable [source: own]

From the histogram, it is easy to understand that the variable Purchase Amount has an almost normal distribution. For getting a better picture, skewness and kurtosis are used. Skewness is a measure of the asymmetry of the probability distribution of a random variable about its mean. () If it is between 0.5 and 1, it means that the distribution is moderately skewed.



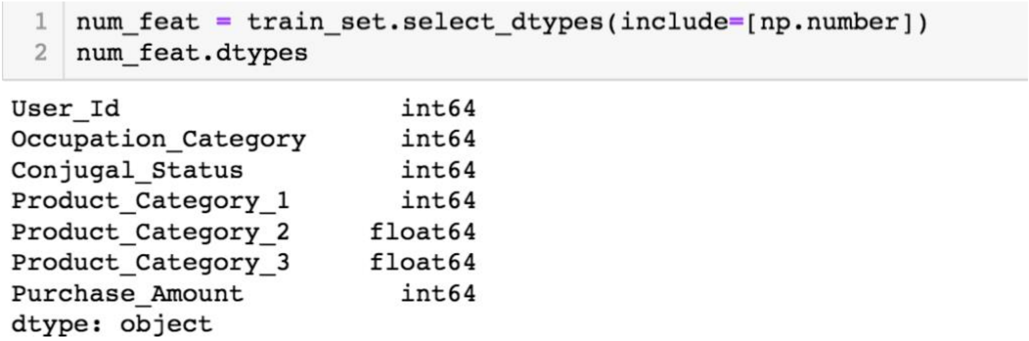
```
print ("Skew:", train_set.Purchase_Amount.skew())
print("Kurtosis: %f" % train_set.Purchase_Amount.kurt())
```

```
Skew: 0.626337134894712
Kurtosis: -0.339678
```

Figure 22 Skewness and Kurtosis of the Distribution of the Dependent Variable [source: own]

From the result, it can be concluded that the distribution is moderately skewed. From the described method above, the mean of the Purchase Amount is 9333, which measures the center of the distribution of the data.

The numerical variables are as shown below:



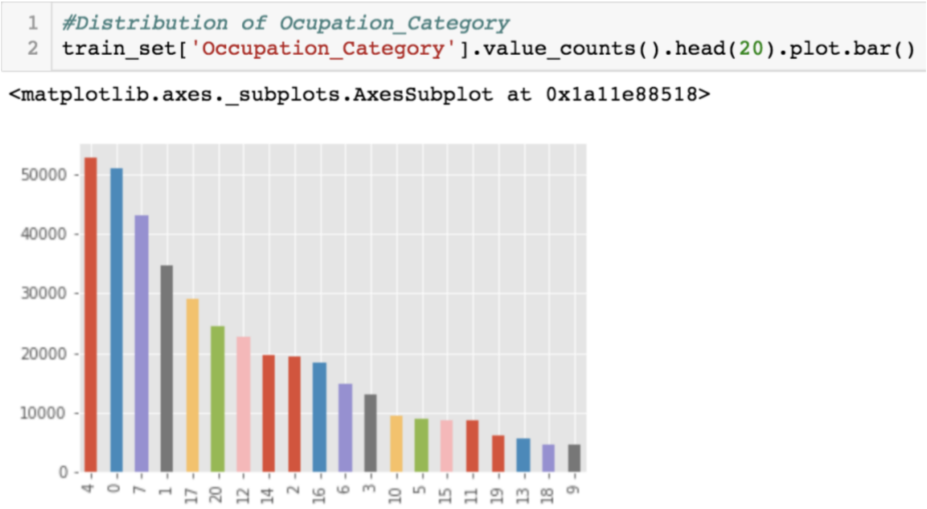
```
num_feat = train_set.select_dtypes(include=[np.number])
num_feat.dtypes
```

```
User_Id                    int64
Occupation_Category        int64
Conjugal_Status            int64
Product_Category_1         int64
Product_Category_2       float64
Product_Category_3       float64
Purchase_Amount            int64
dtype: object
```

Figure 23 Numerical Variables of the dataset [source: own]

Variable: Occupation Category

This variable is not expressed in real values but rather in numerical ones, where each of them identifies one occupation type. Unfortunately, there is no information provided about which number corresponds to which occupation, so the analysis will not tell much. However, it is still possible to see the distribution of the variable.



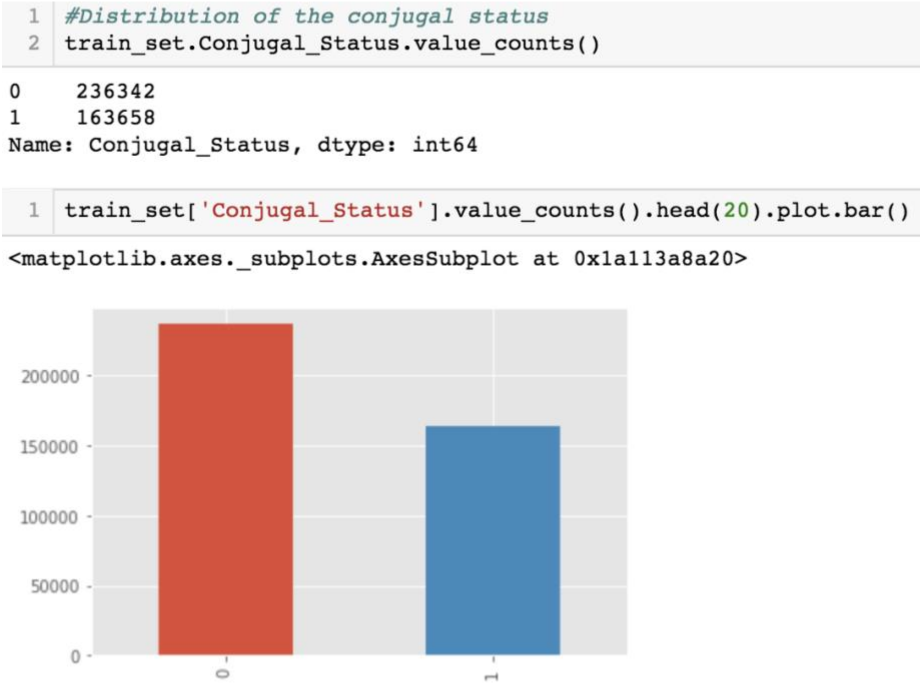
```
#Distribution of Ocupation_Category
train_set['Occupation_Category'].value_counts().head(20).plot.bar()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a11e88518>
```

Figure 24 Distribution of Occupation_Category [source: own]

From the graph, category 4 and 0 have the highest number of customers. This doesn't necessarily mean that the customers from these occupation categories have done the highest amount of purchases. To define such correlation, the bivariate analysis is needed between the two variables. Even though the specific occupations are not present in the dataset, it still helps analysts to gain insight and to take better decisions.

Variable: Conjugal status



```
#Distribution of the conjugal status
train_set.Conjugal_Status.value_counts()
```

```
0    236342
1    163658
Name: Conjugal_Status, dtype: int64
```

```
train_set['Conjugal_Status'].value_counts().head(20).plot.bar()
```

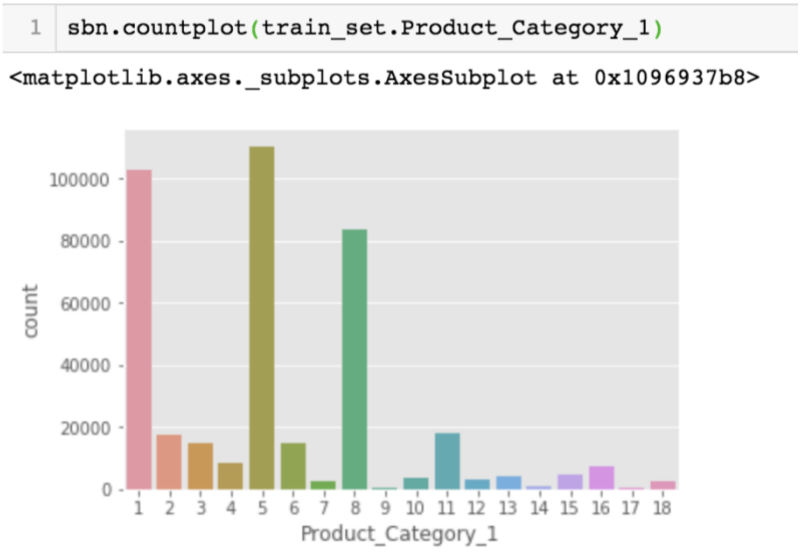
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a113a8a20>
```

Figure 25 Distribution of Conjugal_Status [source: own]

From the histogram, the dataset has 30% more single customers (with 0 status) than married ones. Even though there are different reports stating controversies related to this metric when compared to the dependent variable, single people, especially single women tend to spend more, as Mark Dolliver states in one of his articles: The predisposition for expenditures in clothing is 43% of the singles vs 28% of the married. (Dolliver, 2009) However, for this specific dataset, the bivariate analysis will be done in the following section.

From the chart below, the product category 1 has the highest count of purchases on the Product with code 5. As it is not possible to retrieve the product category name and the product name, this distribution doesn’t tell much for further analysis. Therefore, the other product categories won’t be analyzed.

Variable: Product_Category_1



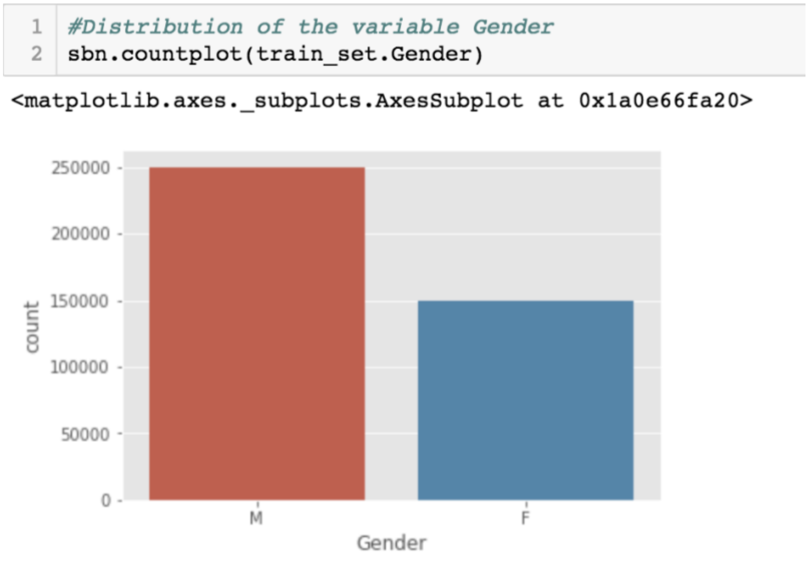
```
sbn.countplot(train_set.Product_Category_1)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1096937b8>
```

Figure 26 Product_Category_1 distribution [source: own]

Variable: Gender

From the chart below, 62% of the customers are male. This shows the distribution of gender independently from the purchase amount. In the next section, this variable will be part of the bivariate analysis as well, where more insights will be gained on who spends more.

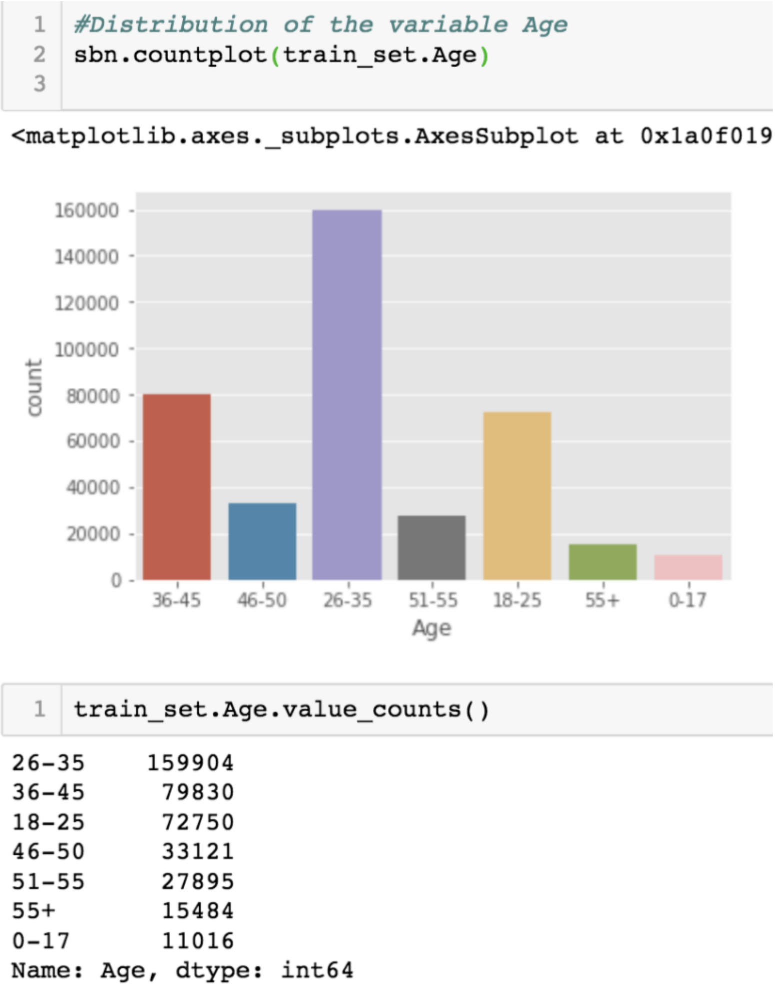


```
#Distribution of the variable Gender
sbn.countplot(train_set.Gender)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a0e66fa20>
```

Figure 27 Gender split in the dataset

Variable Age



```
#Distribution of the variable Age
sbn.countplot(train_set.Age)

```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a0f019
```

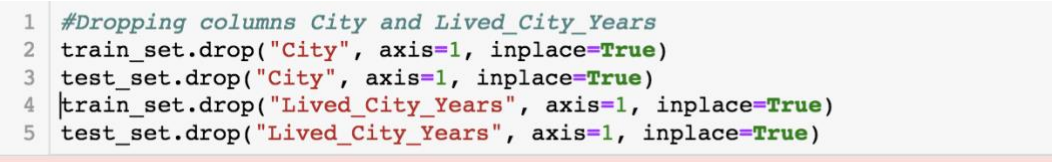
```
train_set.Age.value_counts()
```

```
26-35     159904
36-45      79830
18-25      72750
46-50      33121
51-55      27895
55+        15484
0-17       11016
Name: Age, dtype: int64
```

Figure 28 Distribution of the variable age [source: own]

From the histogram, the age category which has more customers is between 26-35. If a merge of the categories would be done on macro-categories, it can be concluded that the age category from 18 – 35, is the one with more buyers in Category 1.

The variables City and Lived_City_Years won't be taken into consideration as the dataset is focused on e-commerce purchases and none of these purchases has been done onsite. Therefore, they will be dropped as below.

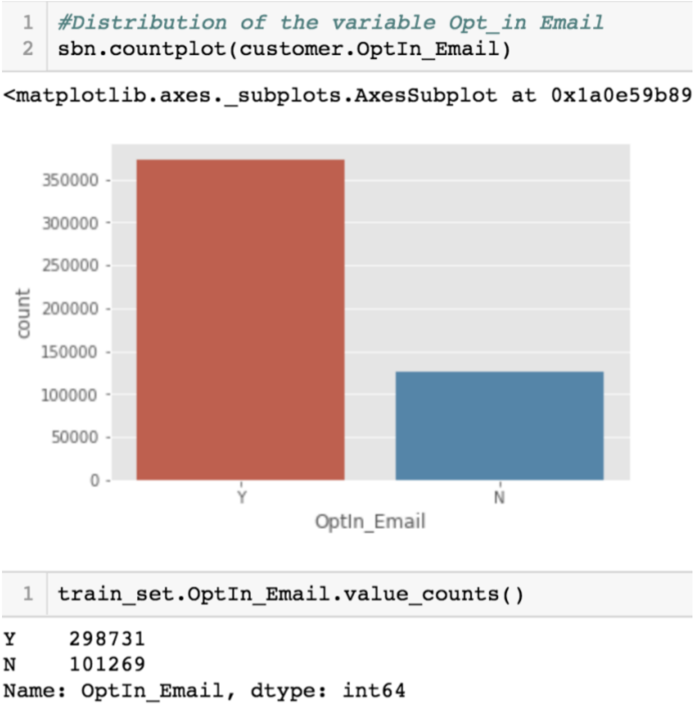


```
#Dropping columns City and Lived_City_Years
train_set.drop("City", axis=1, inplace=True)
test_set.drop("City", axis=1, inplace=True)
train_set.drop("Lived_City_Years", axis=1, inplace=True)
test_set.drop("Lived_City_Years", axis=1, inplace=True)
```

Figure 29 Dropping irrelevant columns [source: own]

Variable OptIn_Email

This variable gives the information whether the buyer has opted in for marketing email communication or not. From the histogram, 66% of the customers are opted it, which means they are interested in receiving marketing campaigns.



```
#Distribution of the variable Opt_in Email
sbn.countplot(customer.OptIn_Email)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a0e59b89
```

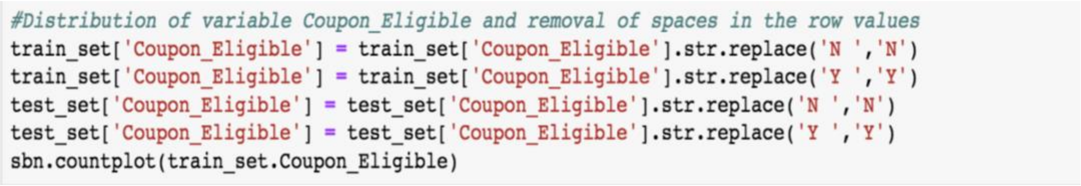
```
train_set.OptIn_Email.value_counts()
```

```
Y    298731
N    101269
Name: OptIn_Email, dtype: int64
```

Figure 30 OptIn_Email distribution [source: own]

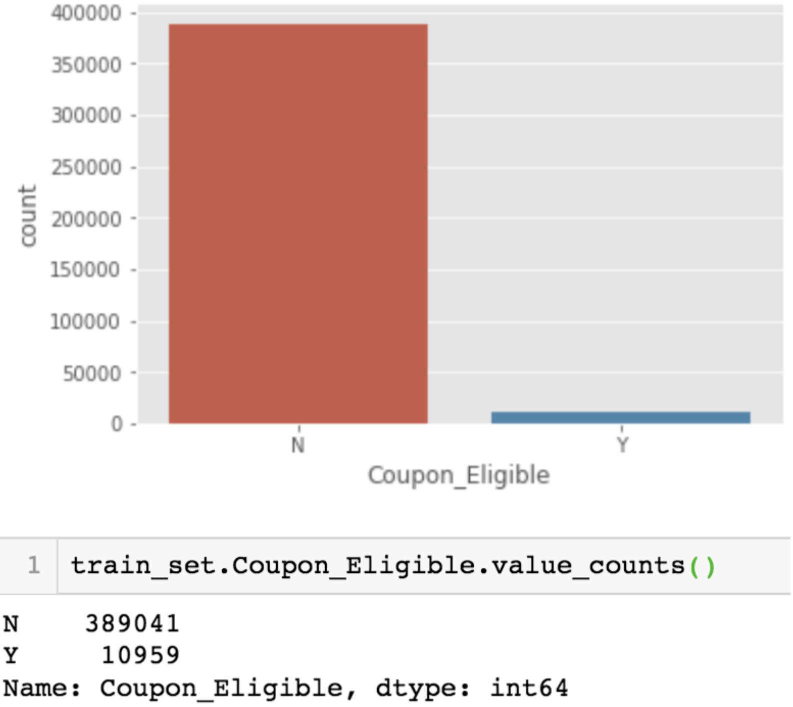
Variable Coupon_Eligible

This variable shows if the customer is eligible for a coupon to redeem. Whether the customer is eligible or not, is selected based on conditions like how much the buyer has purchased, if they are repeated customers or not and other conditions on which marketing campaigns are based. Before plotting the variable, a few data cleaning has been applied as empty spaces were present in the values of it, making more than 2 categories of values (Y, N). Only 2% of the train set is eligible for a coupon, as stated from the numbers above. From this, a conclusion can be drawn even for the variable Item_To_Buy_With_Coupon and Saved_Amount as they are related variables. Whenever a customer is eligible to a coupon, the amount that they can save by redeeming it and the specific product on which they can apply this coupon is present. In the bivariate statistical analysis, they will be checked against the dependent variable.



```
#Distribution of variable Coupon_Eligible and removal of spaces in the row values
train_set['Coupon_Eligible'] = train_set['Coupon_Eligible'].str.replace('N ','N')
train_set['Coupon_Eligible'] = train_set['Coupon_Eligible'].str.replace('Y ','Y')
test_set['Coupon_Eligible'] = test_set['Coupon_Eligible'].str.replace('N ','N')
test_set['Coupon_Eligible'] = test_set['Coupon_Eligible'].str.replace('Y ','Y')
sbn.countplot(train_set.Coupon_Eligible)
```

Figure 31 Data Cleaning for variable Coupon_Eligible [source: own]



```
train_set.Coupon_Eligible.value_counts()
```

```
N    389041
Y     10959
Name: Coupon_Eligible, dtype: int64
```

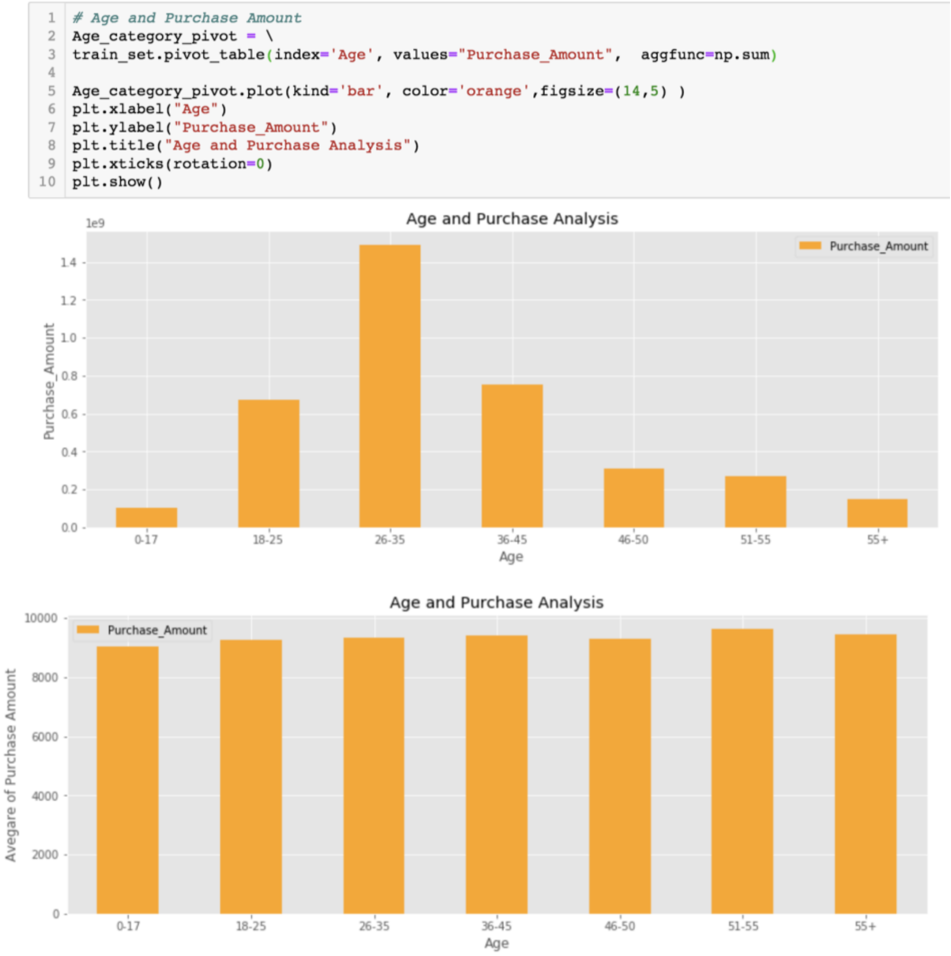
Figure 32 Distribution of the variable Coupon_Eligible

4.2.2 Bivariate statistical analysis

After analyzing each variable independently, it is very important to understand the relationship of the target variable with the independent ones. This is done through the bivariate analysis. The Purchase Amount variable will be on a log scale, due to the high differences among the values.

Age and Purchase Amount

From the bar chart below, it is obvious, that the age category with the highest number of purchases is 26-35. Also, the total amount of purchase spent in accordance with the number of purchases per age category, belongs to people between 26-35. If the same relationship would be plotted in an average of the total amount spent per age category, curiously, that would be higher for buyers who are 50 years or older.

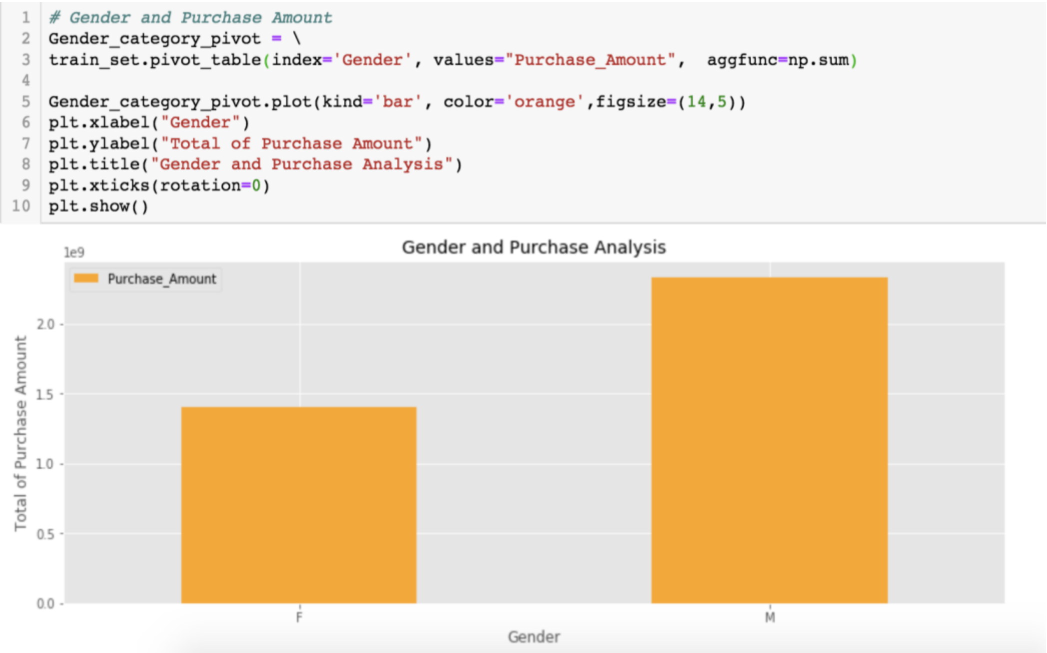


```
# Age and Purchase Amount
Age_category_pivot = \
train_set.pivot_table(index='Age', values="Purchase_Amount",  aggfunc=np.sum)

Age_category_pivot.plot(kind='bar', color='orange',figsize=(14,5) )
plt.xlabel("Age")
plt.ylabel("Purchase_Amount")
plt.title("Age and Purchase Analysis")
plt.xticks(rotation=0)
plt.show()
```

Figure 33 Figure 33 Analysis of Age Category and Purchase Amount [source: own]

Gender and Purchase Amount Analysis

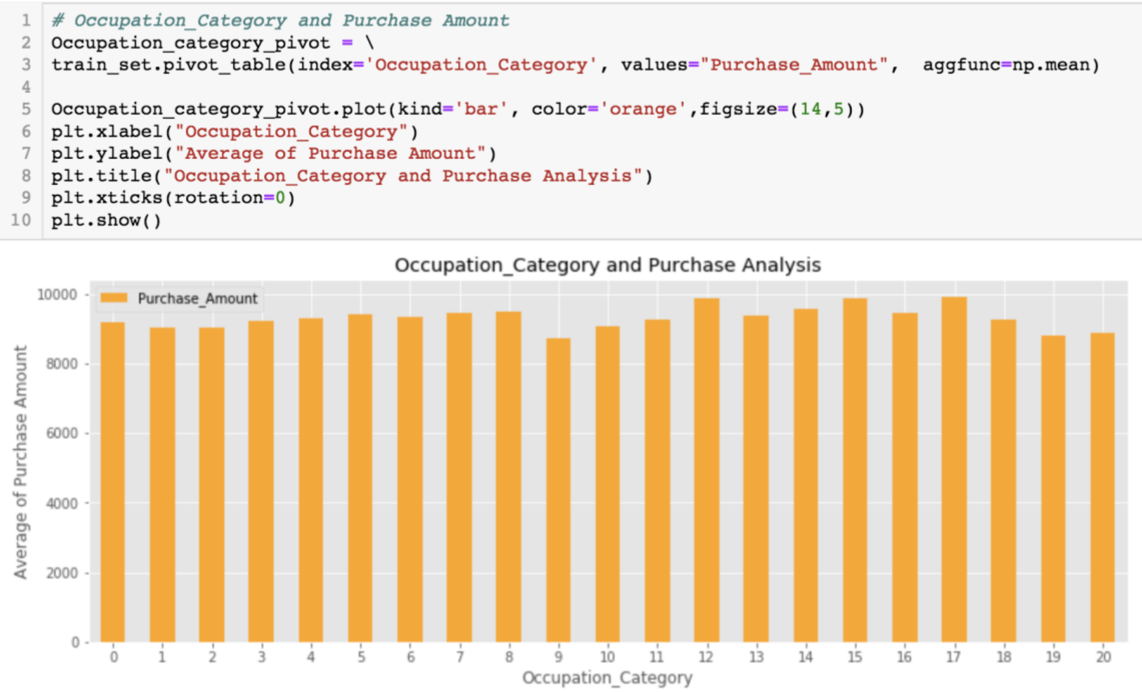
The highest number of buyers in the dataset is male, as analyzed in the univariate distribution of the variable. From the bar chart below, males spend more on a total purchase amount as well as on an average amount, contrary to women. This result is very important for the marketers of the company when creating the marketing campaigns per gender split.



```
# Gender and Purchase Amount
Gender_category_pivot = \
train_set.pivot_table(index='Gender', values="Purchase_Amount",  aggfunc=np.sum)

Gender_category_pivot.plot(kind='bar', color='orange',figsize=(14,5))
plt.xlabel("Gender")
plt.ylabel("Total of Purchase Amount")
plt.title("Gender and Purchase Analysis")
plt.xticks(rotation=0)
plt.show()
```

Figure 34 Total Purchase Amount per Gender [source: own]

Occupation Category and Purchase Amount Analysis

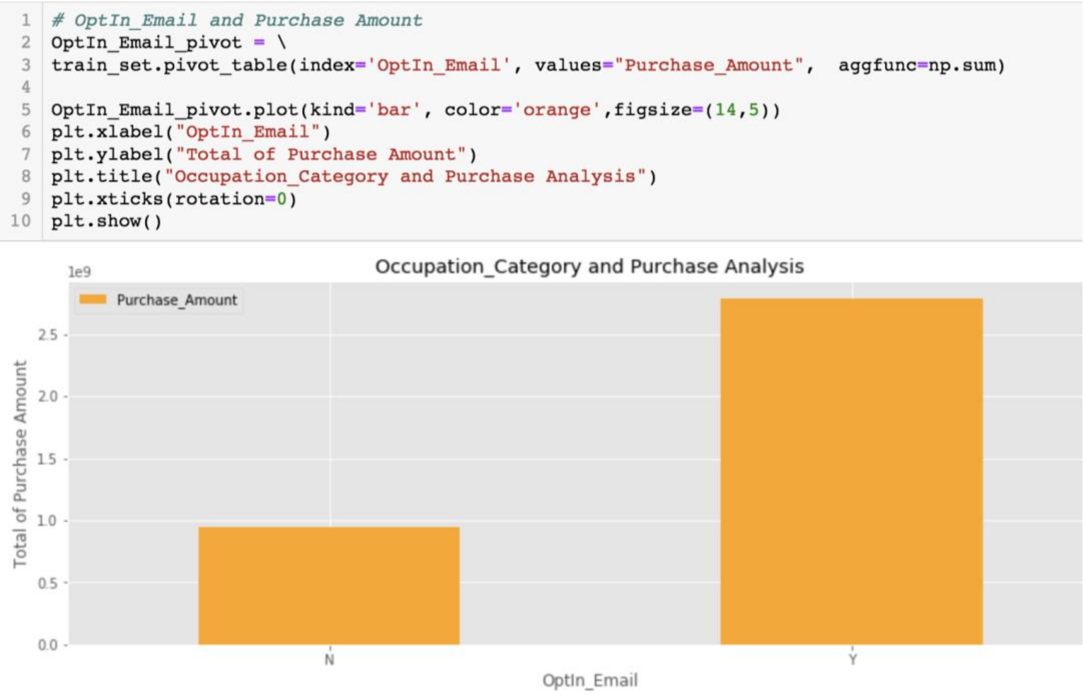


```
# Occupation_Category and Purchase Amount
Occupation_category_pivot = \
train_set.pivot_table(index='Occupation_Category', values="Purchase_Amount",  aggfunc=np.mean)

Occupation_category_pivot.plot(kind='bar', color='orange',figsize=(14,5))
plt.xlabel("Occupation_Category")
plt.ylabel("Average of Purchase Amount")
plt.title("Occupation_Category and Purchase Analysis")
plt.xticks(rotation=0)
plt.show()
```

Figure 35 Average expenditure from each occupation category [source: own]

The amount that customers spend on an average from each occupation category, is pretty much the same, even though it is noticeable that the occupation coded 17 has the highest average purchase. From this thesis analysis, there is not much of insight that can be gained, but for the analysts and marketer of the ABCommerce who have access on the specific category nominations, this can help identify the most important professions to be targeted.

OptIn Email and Purchase Amount Analysis



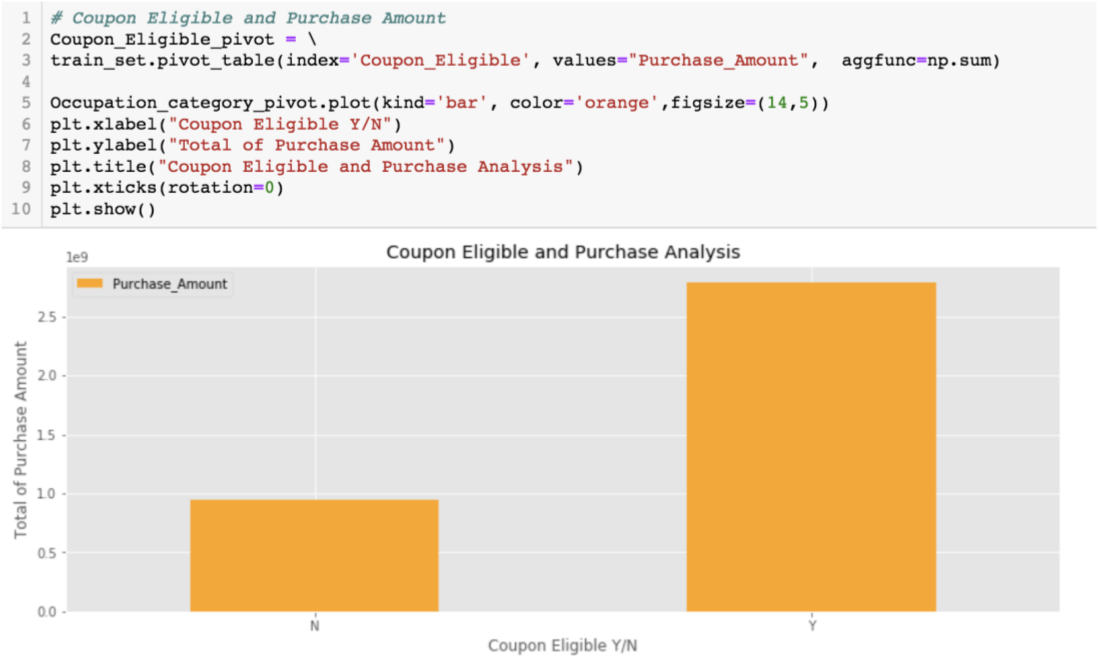
```
# OptIn_Email and Purchase Amount
OptIn_Email_pivot = \
train_set.pivot_table(index='OptIn_Email', values="Purchase_Amount",  aggfunc=np.sum)

OptIn_Email_pivot.plot(kind='bar', color='orange',figsize=(14,5))
plt.xlabel("OptIn_Email")
plt.ylabel("Total of Purchase Amount")
plt.title("Occupation_Category and Purchase Analysis")
plt.xticks(rotation=0)
plt.show()
```

Figure 36 Analysis between OptIn_Email and Purchase Amount

Customers who are opted in to marketing email communications, have spent 66% more than buyers who are not opted in. This clearly states the importance of email communication in increasing the overall expenditure level.

Coupon Eligible and Purchase Amount

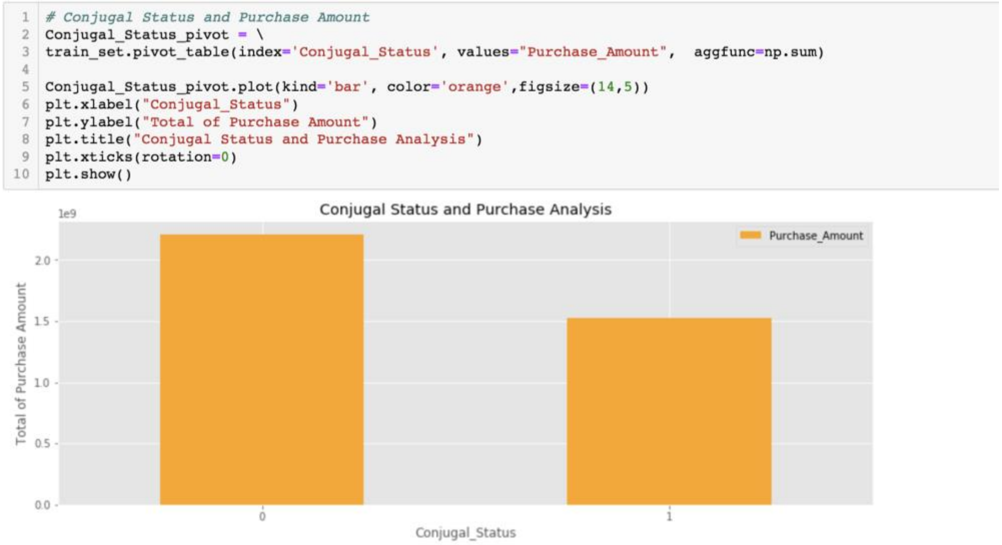
As predicted, the highest amount of purchases, belongs to customers who are eligible to redeem a coupon. This is studied also from the company itself, which provides coupon to high frequent spenders and from which the revenues boost significantly.



```python
# Coupon Eligible and Purchase Amount
Coupon_Eligible_pivot = \
train_set.pivot_table(index='Coupon_Eligible', values="Purchase_Amount",  aggfunc=np.sum)

Occupation_category_pivot.plot(kind='bar', color='orange',figsize=(14,5))
plt.xlabel("Coupon Eligible Y/N")
plt.ylabel("Total of Purchase Amount")
plt.title("Coupon Eligible and Purchase Analysis")
plt.xticks(rotation=0)
plt.show()
```

Figure 37 Coupon Eligible and Purchase Amount Analysis

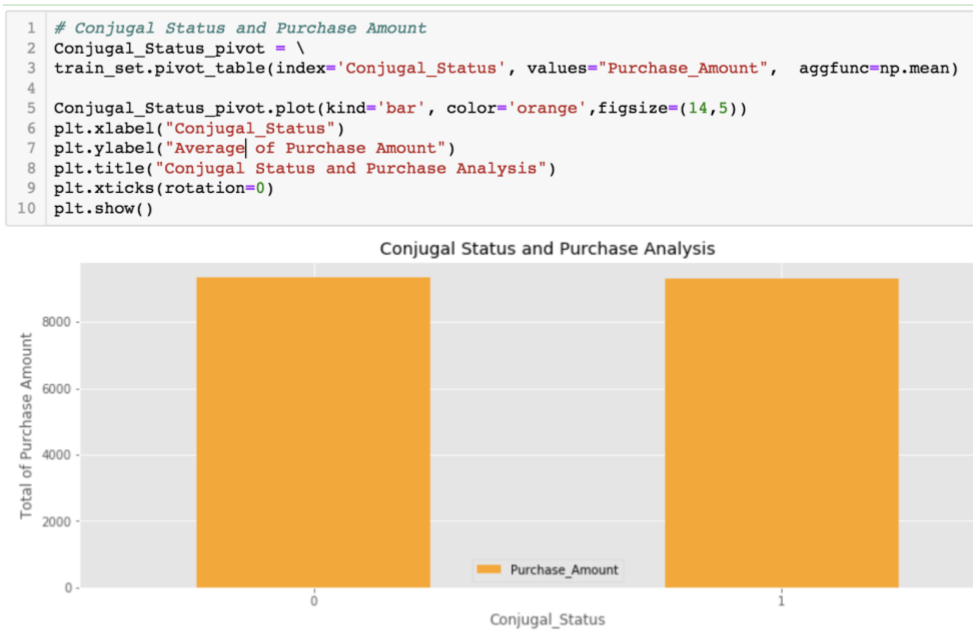
Conjugal Status and Purchase Amount Analysis



```python
# Conjugal Status and Purchase Amount
Conjugal_Status_pivot = \
train_set.pivot_table(index='Conjugal_Status', values="Purchase_Amount",  aggfunc=np.sum)

Conjugal_Status_pivot.plot(kind='bar', color='orange',figsize=(14,5))
plt.xlabel("Conjugal_Status")
plt.ylabel("Total of Purchase Amount")
plt.title("Conjugal Status and Purchase Analysis")
plt.xticks(rotation=0)
plt.show()
```

Figure 38 Conjugal Status and Purchase Amount [source: own]

There are more single customers than married ones in the dataset and in additional single buyers have purchased more in total than the married ones. However, what is interesting, is that on an average scale, individual customers tend to spend the same, no matter their conjugal status. Since this variable has a higher representation on the single category, it makes more sense to see the correlation in an average then as a sum of purchases.



```python
# Conjugal Status and Purchase Amount
Conjugal_Status_pivot = \
train_set.pivot_table(index='Conjugal_Status', values="Purchase_Amount",  aggfunc=np.mean)

Conjugal_Status_pivot.plot(kind='bar', color='orange',figsize=(14,5))
plt.xlabel("Conjugal_Status")
plt.ylabel("Average of Purchase Amount")
plt.title("Conjugal Status and Purchase Analysis")
plt.xticks(rotation=0)
plt.show()
```

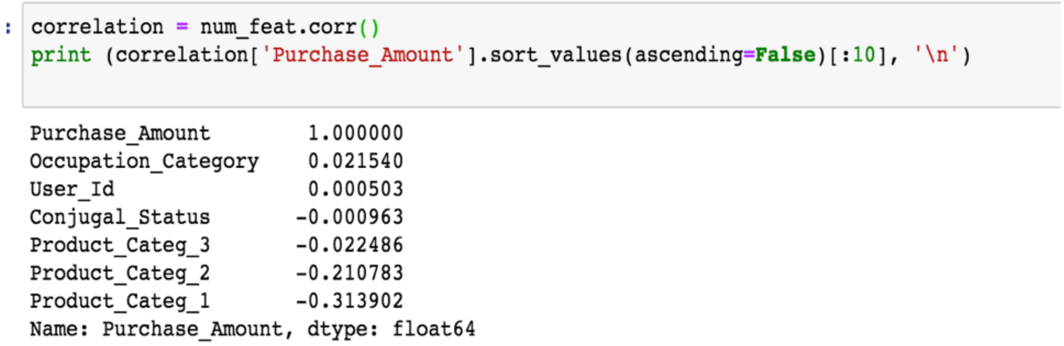
Figure 39 Average of Purchase Amount per Conjugal_Status [source: own]

In the next section, the correlation between the dependent variable and the regressors will be analyzed.

4.2.3 Correlation between the dependent variable and the numerical ones

Correlation analysis measures the degree of relationship between the target variable and one predictor. (Business Statistics, 2004)

From the correlation method core (), the following table is generated:



```python
correlation = num_feat.corr()
print (correlation['Purchase_Amount'].sort_values(ascending=False)[:10], '\n')
```

```
Purchase_Amount        1.000000
Occupation_Category    0.021540
User_Id                0.000503
Conjugal_Status       -0.000963
Product_Categ_3       -0.022486
Product_Categ_2       -0.210783
Product_Categ_1       -0.313902
Name: Purchase_Amount, dtype: float64
```

Figure 40 Correlations with the target variable [source: own]

Oddly, there is no strong correlation between Purchase_Amount and any of the other variables. The correlation is stronger when the coefficient is closer to 1.0. The only statistically significant correlation is with Occupation_Category which makes sense, as people of different occupations tend to spend differently. Also, there is a statistically significant negative relationship between Purchase_Amount and Product_Category_1. This is one of the many challenges when analysing Big Data, what to do when there is no strong correlation among the target variable and any of the independent ones. One of the solutions is to replace the empty values with the mean or median or to remove the variables where more than half of the values are empty.

From the other hand, multicollinearity is the possible relationship amongst predictors. This should not exist as it causes problems to fit the model and also to 'trusting' the statistically important variables. To check the multicollinearity, the correlation matrix is needed.

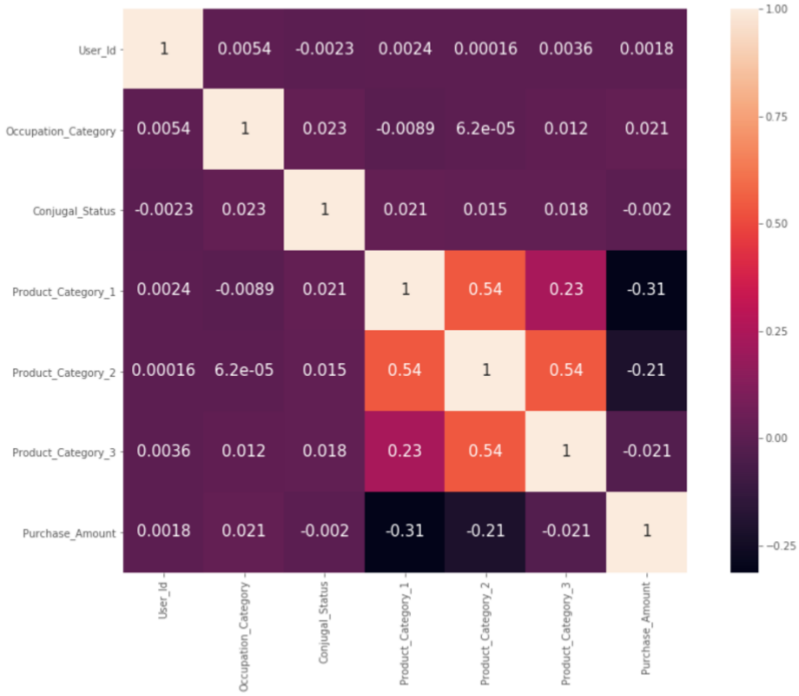


Figure 41 Correlation Matrix between Variables [source: own]

It looks like there is no high multicollinearity among the independent variables apart Product_Category_2 and Product_Category_3 where it seems to be a relatively high correlation. Even between Product_Category_1 and Product_Category_2, there is high multicollinearity. One of the recommended solutions in such cases is to remove one of the variables as they offer redundant information. Both Product_Category_2 and

Product_Category_3 can be removed. Such custom transformations before building the models are applied in the next section.

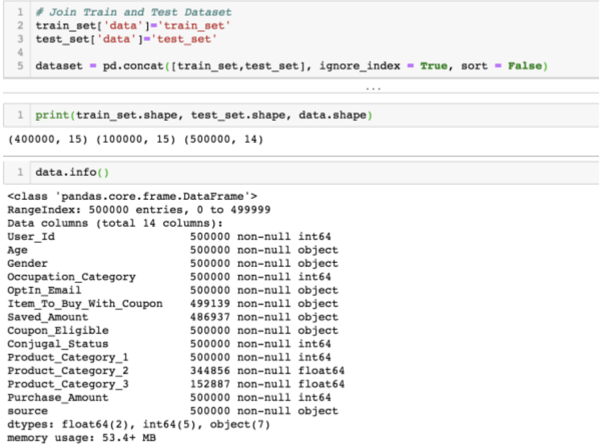
4.3 Model building and training

4.3.1 Custom transformation of the data

Some of the data transformations are already done as part of the EDA. This involved the drop of the variables City and Lived_City_Year as agreed that they are irrelevant for online purchases. Some other recommendations for the other variables would be:

- There are 20 occupations categories which are too many. By checking the frequency, they can be reduced to less.
- Since the variables Gender, OptIn_Email and Coupon_Eligible, have only 2 values, they can be transformed into binary.
- There are empty values on each of the product category fields for which a decision needs to be taken on what to do with them.

Before starting the data cleaning for modeling it, it is a best practice to join the test set with the train set. The only reason to do this, is for not having to repeat the steps again for both datasets. After combining, the new dataframe won't have the City and Lived_City_Years variable which have been already dropped.



```
# Join Train and Test Dataset
train_set['data']='train_set'
test_set['data']='test_set'

dataset = pd.concat([train_set,test_set], ignore_index = True, sort = False)
```

```
print(train_set.shape, test_set.shape, data.shape)
```

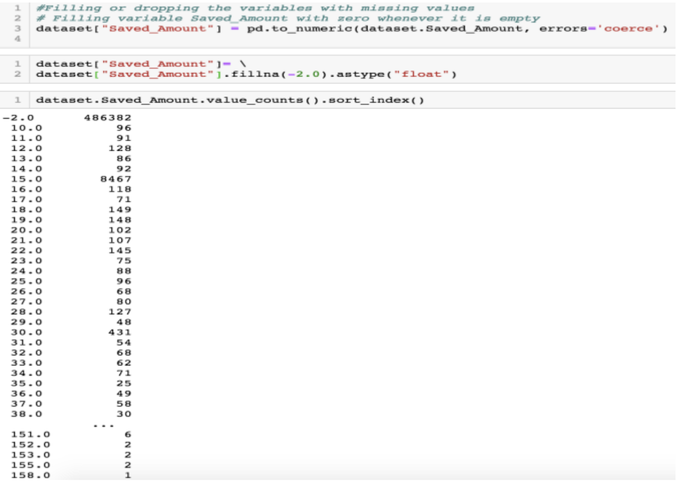
```
(400000, 15) (100000, 15) (500000, 14)
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500000 entries, 0 to 499999
Data columns (total 14 columns):
User_Id                    500000 non-null int64
Age                        500000 non-null object
Gender                     500000 non-null object
Occupation_Category        500000 non-null int64
OptIn_Email                500000 non-null object
Item_To_Buy_With_Coupon    499139 non-null object
Saved_Amount               486937 non-null object
Coupon_Eligible            500000 non-null object
Conjugal_Status            500000 non-null int64
Product_Category_1         500000 non-null int64
Product_Category_2         344856 non-null float64
Product_Category_3         152887 non-null float64
Purchase_Amount            500000 non-null int64
source                     500000 non-null object
dtypes: float64(2), int64(5), object(7)
memory usage: 53.4+ MB
```

Figure 42 New dataset after joining train and test sets [source: own]

As seen from the figure above, there are a few variables which have null values. They are either removed completely as variables or the missing values can be replaced with either zero or the mean of the distribution of that predictor. The first one to be filled with zeros is the Saved_Amount as below:



```
#Filling or dropping the variables with missing values
# Filling variable Saved_Amount with zero whenever it is empty
dataset["Saved_Amount"] = pd.to_numeric(dataset.Saved_Amount, errors='coerce')

```

```
dataset["Saved_Amount"]= \
dataset["Saved_Amount"].fillna(-2.0).astype("float")
```

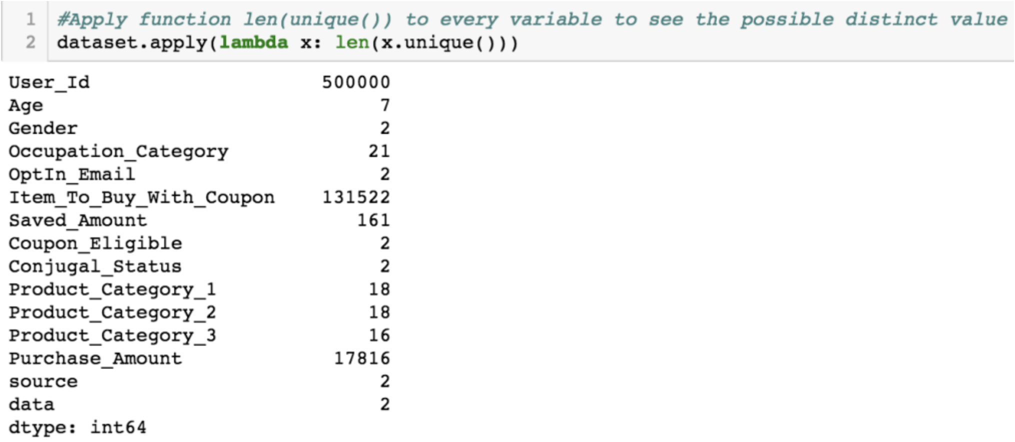
```
dataset.Saved_Amount.value_counts().sort_index()
```

```
-2.0      486382
 10.0         96
 11.0         91
 12.0        128
 13.0         86
 14.0         92
 15.0       8467
 16.0        118
 17.0         71
 18.0        149
 19.0        148
 20.0        102
 21.0        107
 22.0        145
 23.0         75
 24.0         88
 25.0         96
 26.0         68
 27.0         80
 28.0        127
 29.0         48
 30.0        431
 31.0         54
 32.0         68
 33.0         62
 34.0         71
 35.0         25
 36.0         49
 37.0         58
 38.0         30
              ...
 151.0         6
 152.0         2
 153.0         2
 155.0         2
 158.0         1
```

Figure 43 Imputing Values in Saved_Amount variable [source: own]

The same can be done with the other variables such as Product Category 2 and Product Category 3. The figures won't be included here, but the commands have been applied in Python.

As for the other categorical variables, they need to be transformed from their existing types to different ones to better fit the techniques afterwards. Below are the counts of distinct values for each variable present in the dataset.



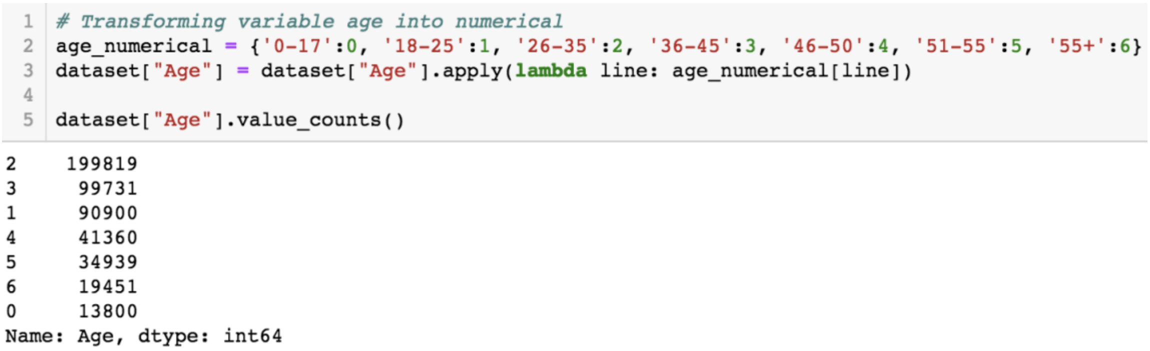
```
#Apply function len(unique()) to every variable to see the possible distinct value
dataset.apply(lambda x: len(x.unique()))
```

```
User_Id                     500000
Age                              7
Gender                           2
Occupation_Category             21
OptIn_Email                      2
Item_To_Buy_With_Coupon     131522
Saved_Amount                   161
Coupon_Eligible                  2
Conjugal_Status                  2
Product_Category_1              18
Product_Category_2              18
Product_Category_3              16
Purchase_Amount              17816
source                           2
data                             2
dtype: int64
```

Figure 44 Checking the data values for each variable before conversion [source: own]

Age variable

The variable age should be treated as numerical, since it represents a number, and having values such as 55+ makes it more difficult for the predictor to be part of the model.



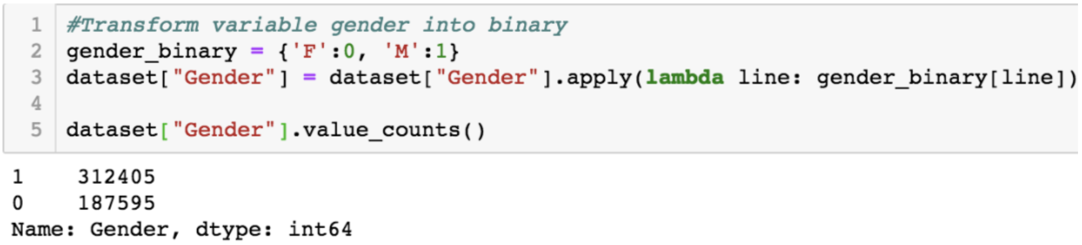
```
# Transforming variable age into numerical
age_numerical = {'0-17':0, '18-25':1, '26-35':2, '36-45':3, '46-50':4, '51-55':5, '55+':6}
dataset["Age"] = dataset["Age"].apply(lambda line: age_numerical[line])

dataset["Age"].value_counts()
```

```
2    199819
3     99731
1     90900
4     41360
5     34939
6     19451
0     13800
Name: Age, dtype: int64
```

Figure 45 The transformation of variable age into numerical [source: own]

Gender variable

The variable gender should be converted into binary for feature engineering purposes, otherwise it won’t be possible to include it in the linear regression model. The same for Coupon_Eligible and OptIn_Email.



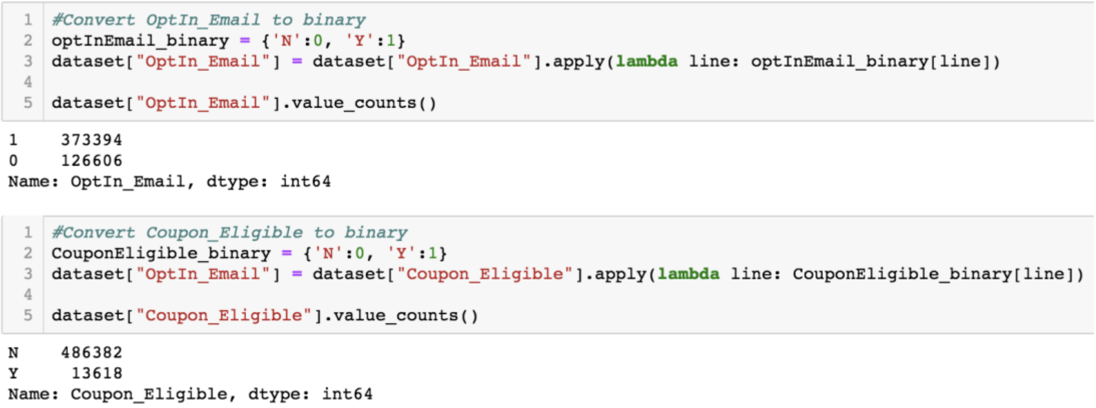
```
#Transform variable gender into binary
gender_binary = {'F':0, 'M':1}
dataset["Gender"] = dataset["Gender"].apply(lambda line: gender_binary[line])

dataset["Gender"].value_counts()
```

```
1    312405
0    187595
Name: Gender, dtype: int64
```

Figure 46 Transformation of variable gender [source: own]

OptIn_Email and Coupon_Eligible variables

The same procedure and the same code that was used for transforming the variable gender and age, is used for variable OptIn_Email to change it into a binary variable.



```
#Convert OptIn_Email to binary
optInEmail_binary = {'N':0, 'Y':1}
dataset["OptIn_Email"] = dataset["OptIn_Email"].apply(lambda line: optInEmail_binary[line])

dataset["OptIn_Email"].value_counts()
```

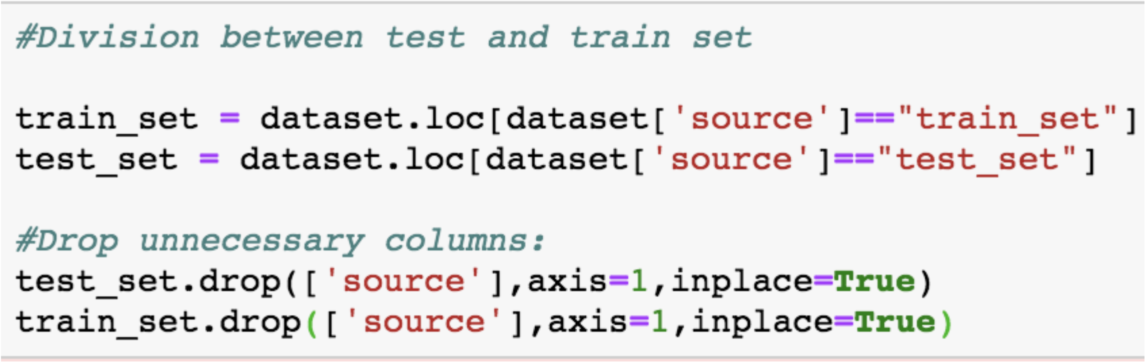
```
1    373394
0    126606
Name: OptIn_Email, dtype: int64
```

```
#Convert Coupon_Eligible to binary
CouponEligible_binary = {'N':0, 'Y':1}
dataset["OptIn_Email"] = dataset["Coupon_Eligible"].apply(lambda line: CouponEligible_binary[line])

dataset["Coupon_Eligible"].value_counts()
```

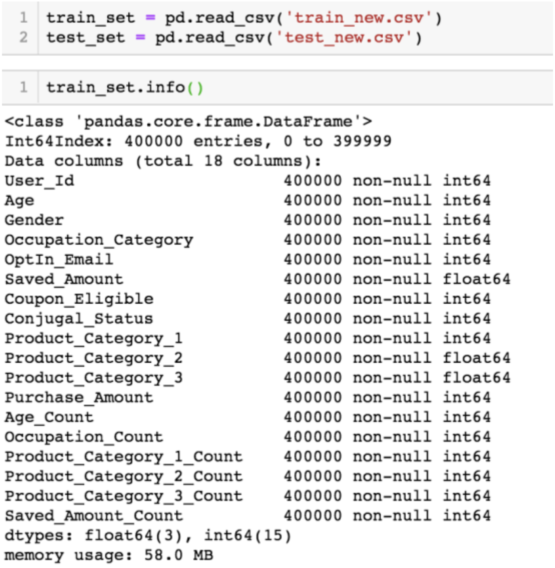
```
N    486382
Y     13618
Name: Coupon_Eligible, dtype: int64
```

Figure 47 Transformation of OptIn_Email and Coupon_Eligible variables into binary ones

All these transformations can be done also automatically through the transformation pipelines found in Sci-Kit learn package, which helps with these automatic and sequential transformations. After all the necessary changes has been applied, the dataset will be split again into test set and train set for training the models.



```
#Division between test and train set

train_set = dataset.loc[dataset['source']=="train_set"]
test_set = dataset.loc[dataset['source']=="test_set"]

#Drop unnecessary columns:
test_set.drop(['source'],axis=1,inplace=True)
train_set.drop(['source'],axis=1,inplace=True)
```



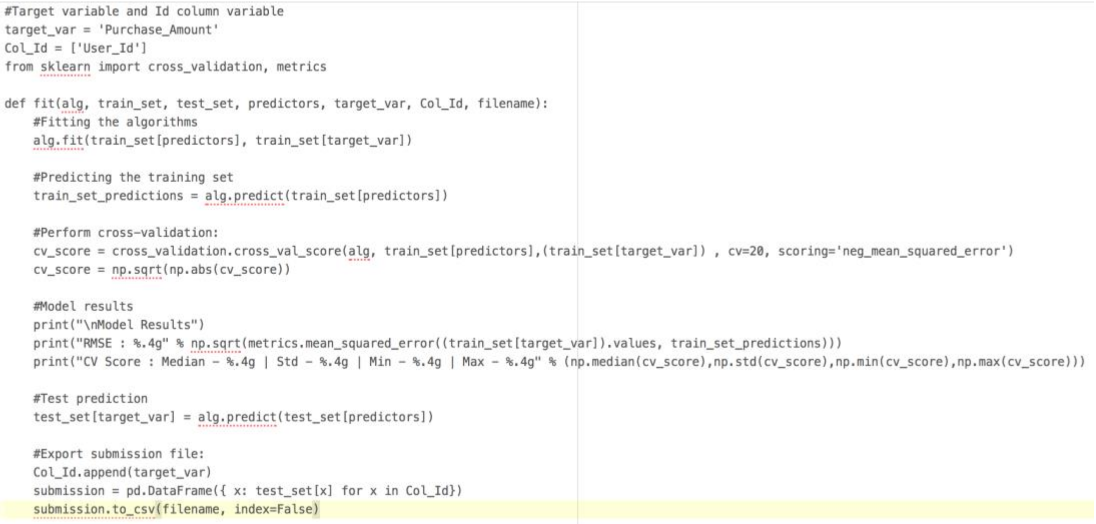
```
train_set = pd.read_csv('train_new.csv')
test_set = pd.read_csv('test_new.csv')
```

```
train_set.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 400000 entries, 0 to 399999
Data columns (total 18 columns):
User_Id                        400000 non-null int64
Age                            400000 non-null int64
Gender                         400000 non-null int64
Occupation_Category            400000 non-null int64
OptIn_Email                    400000 non-null int64
Saved_Amount                   400000 non-null float64
Coupon_Eligible                400000 non-null int64
Conjugal_Status                400000 non-null int64
Product_Category_1             400000 non-null int64
Product_Category_2             400000 non-null float64
Product_Category_3             400000 non-null float64
Purchase_Amount                400000 non-null int64
Age_Count                      400000 non-null int64
Occupation_Count               400000 non-null int64
Product_Category_1_Count       400000 non-null int64
Product_Category_2_Count       400000 non-null int64
Product_Category_3_Count       400000 non-null int64
Saved_Amount_Count             400000 non-null int64
dtypes: float64(3), int64(15)
memory usage: 58.0 MB
```

Figure 48 New train set after all the necessary transformation [source: own]

Now that the problem has been framed, the data has been explored, analyzed, prepared and cleaned, it is ready for Machine Learning algorithms. Since there will be 4 models which will run and generate the performance measure, it is a best practice to create a function which takes the data as an input and uses it to create the model and to perform the cross-validation. Such function is created in the Python code below:



```
#Target variable and Id column variable
target_var = 'Purchase_Amount'
Col_Id = ['User_Id']
from sklearn import cross_validation, metrics

def fit(alg, train_set, test_set, predictors, target_var, Col_Id, filename):
    #Fitting the algorithms
    alg.fit(train_set[predictors], train_set[target_var])

    #Predicting the training set
    train_set_predictions = alg.predict(train_set[predictors])

    #Perform cross-validation:
    cv_score = cross_validation.cross_val_score(alg, train_set[predictors],(train_set[target_var]) , cv=20, scoring='neg_mean_squared_error')
    cv_score = np.sqrt(np.abs(cv_score))

    #Model results
    print("\nModel Results")
    print("RMSE : %.4g" % np.sqrt(metrics.mean_squared_error((train_set[target_var]).values, train_set_predictions)))
    print("CV Score : Median - %.4g | Std - %.4g | Min - %.4g | Max - %.4g" % (np.median(cv_score),np.std(cv_score),np.min(cv_score),np.max(cv_score)))

    #Test prediction
    test_set[target_var] = alg.predict(test_set[predictors])

    #Export submission file:
    Col_Id.append(target_var)
    submission = pd.DataFrame({ x: test_set[x] for x in Col_Id})
    submission.to_csv(filename, index=False)
```

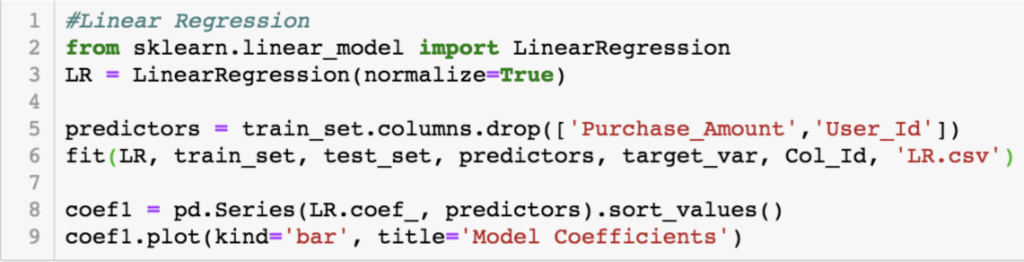
Figure 49 Generic function for model building and cross-validation [source: own]

5. Results and Discussion

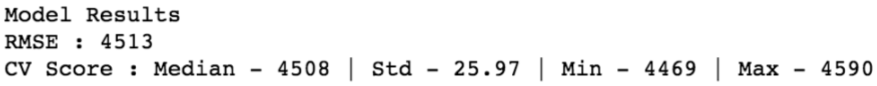
5.1 Model building and evaluation

Linear regression

In order to measure the RMSE on the whole training set, the Scikit Learn function will be built as below:



```
#Linear Regression
from sklearn.linear_model import LinearRegression
LR = LinearRegression(normalize=True)

predictors = train_set.columns.drop(['Purchase_Amount','User_Id'])
fit(LR, train_set, test_set, predictors, target_var, Col_Id, 'LR.csv')

coef1 = pd.Series(LR.coef_, predictors).sort_values()
coef1.plot(kind='bar', title='Model Coefficients')
```



```
Model Results
RMSE : 4513
CV Score : Median - 4508 | Std - 25.97 | Min - 4469 | Max - 4590
```

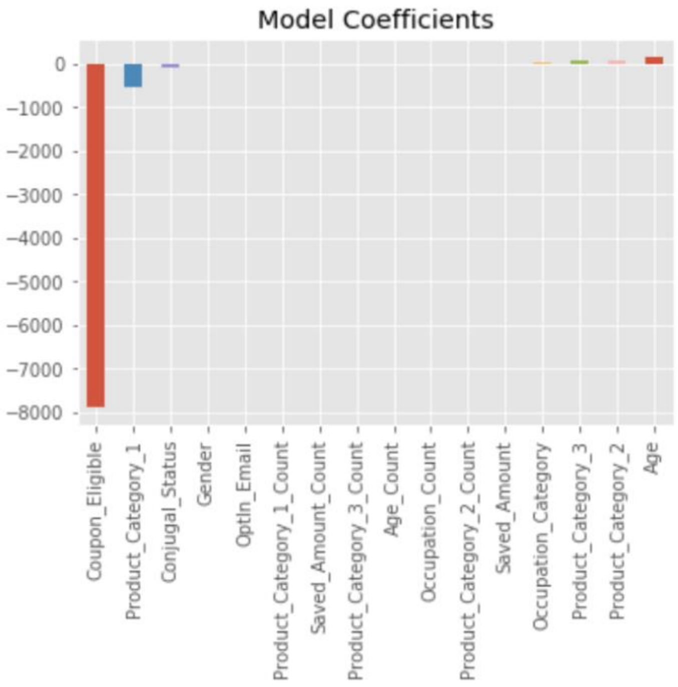
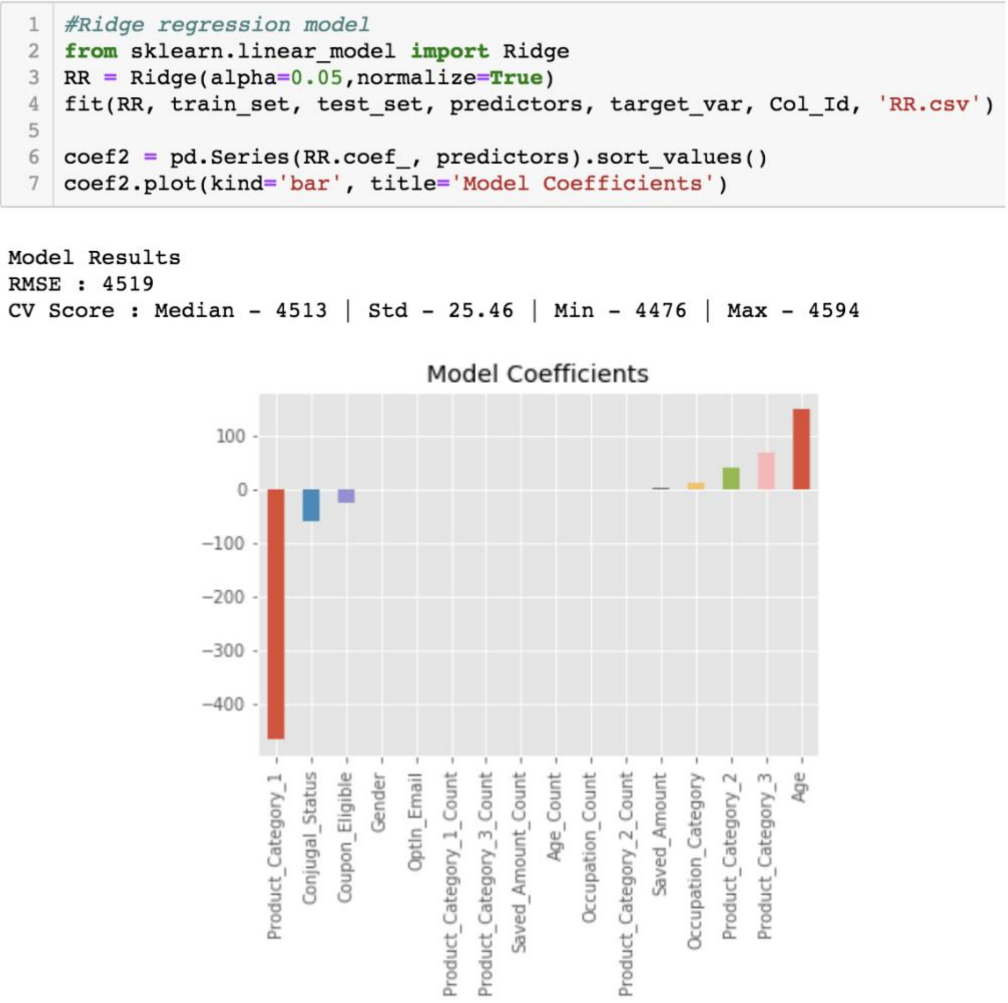


Figure 50 Linear Regression Performance Measure Coefficient [source: own]

From the results, it looks the prediction is quite accurate. The median of the purchase amount is 4590 and the RSME has scored 4513, which is a very satisfying prediction error. However, this has to be compared with the other models, to be able to come to a conclusion on which model is the best one for predicting such variable.

Ridge Regression model

As mentioned in the literature review, ridge regression is just a variation of the linear regression mode used in cases of predictors with high multicollinearity. From the multicollinearity checking, it is high only between Product_Category_1 and Product_Category_2. In such cases, theoretically the standard error should be reduced. This is to be seen in the model application.



```
#Ridge regression model
from sklearn.linear_model import Ridge
RR = Ridge(alpha=0.05,normalize=True)
fit(RR, train_set, test_set, predictors, target_var, Col_Id, 'RR.csv')

coef2 = pd.Series(RR.coef_, predictors).sort_values()
coef2.plot(kind='bar', title='Model Coefficients')
```

```
Model Results
RMSE : 4519
CV Score : Median - 4513 | Std - 25.46 | Min - 4476 | Max - 4594
```

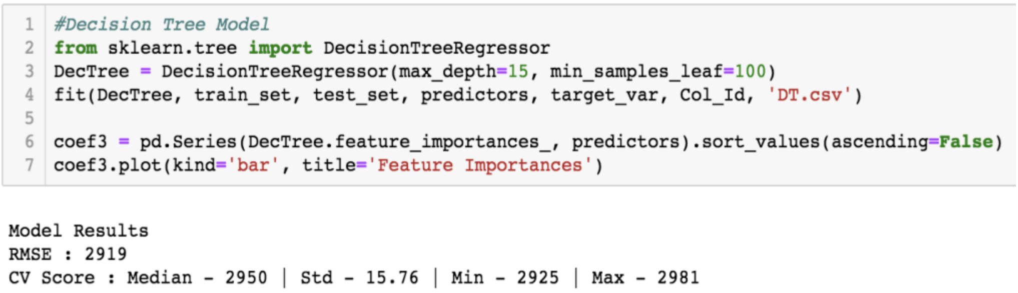
Figure 51 Ridge Regression Model [source: own]

Surprisingly, the ridge regression doesn’t perform better than the linear regression, as its RMSE is higher than in the first case. This can be explained with the non-presence of high multicollinearity amongst independent variables.

Decision Tree model

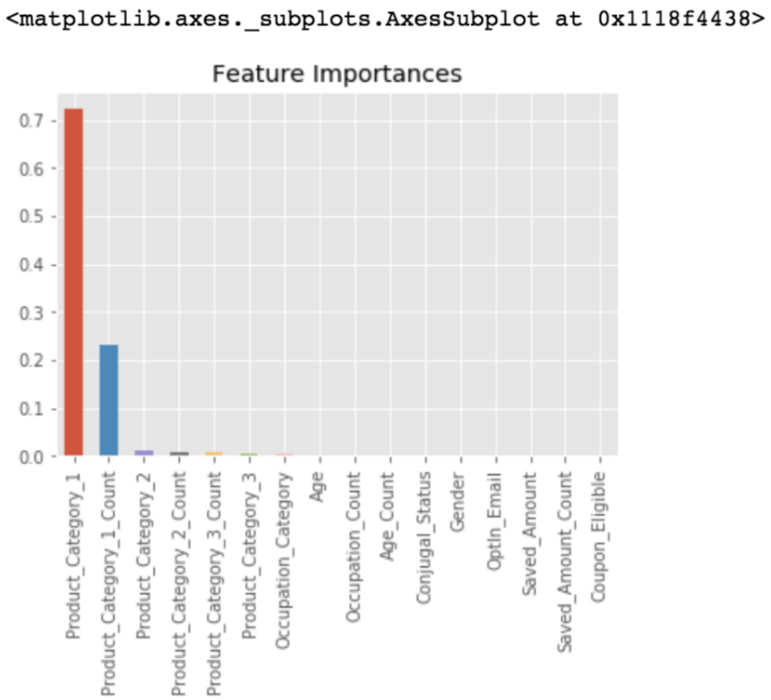
There are different ways on how to build the decision tree model. One way is to split the training test into smaller training sets and a validation set through the function train_test_split. In this way, the model can be trained against smaller training sets and from the other side these can be trained against the validation set. Another way, it to use the SciKit -Learn cross-validation feature. From this feature shown above, the training set is being split into 20 distinct subsets called folds then it trains and evaluates the decision tree model 20 times, picking a different fold for evaluation every time and training on the other 19 folds. (Géron, 2017)

From the figure below, the RSME of the decision tree has the best score so far as it has the lowest one. However, there is still Random Forrest Model which will be checked in the same way.



```
#Decision Tree Model
from sklearn.tree import DecisionTreeRegressor
DecTree = DecisionTreeRegressor(max_depth=15, min_samples_leaf=100)
fit(DecTree, train_set, test_set, predictors, target_var, Col_Id, 'DT.csv')

coef3 = pd.Series(DecTree.feature_importances_, predictors).sort_values(ascending=False)
coef3.plot(kind='bar', title='Feature Importances')
```

```
Model Results
RMSE : 2919
CV Score : Median - 2950 | Std - 15.76 | Min - 2925 | Max - 2981
```

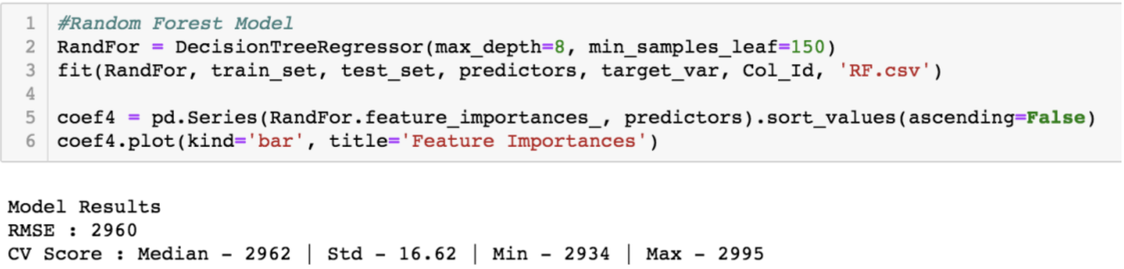


```
<matplotlib.axes._subplots.AxesSubplot at 0x1118f4438>
```

Figure 52 Decision Tree Performance Measure [source: own]

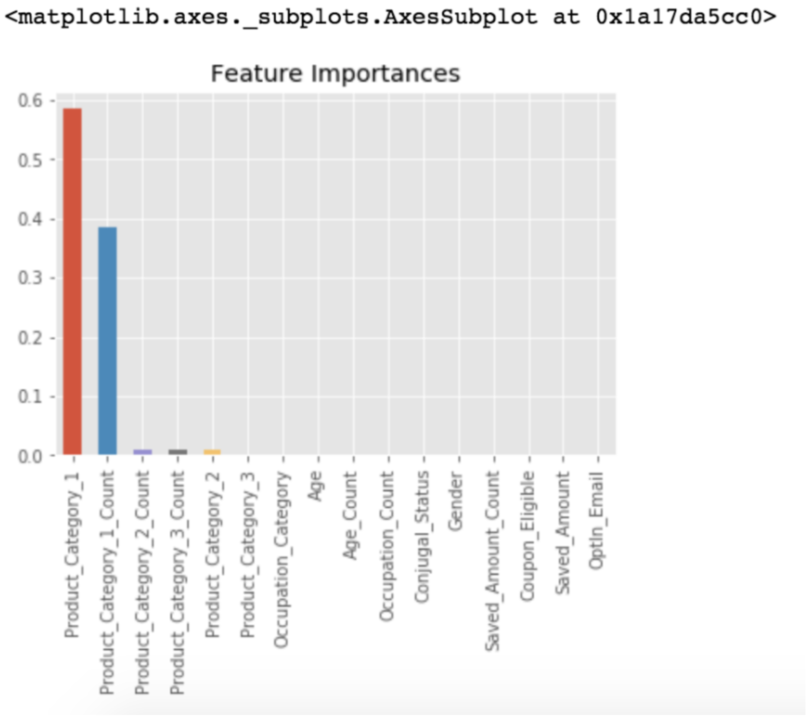
Random Forest Model

Random forest is similar with the decision tree, with the difference that it works by training many decision trees on random subsets of the features and after it calculated the average of these prediction. In this way, a model on top of another model is built and in Machine Learning this is called Ensemble Learning. The code applied will still be the same, so below is the performance measure through RMSE.



```python
#Random Forest Model
RandFor = DecisionTreeRegressor(max_depth=8, min_samples_leaf=150)
fit(RandFor, train_set, test_set, predictors, target_var, Col_Id, 'RF.csv')

coef4 = pd.Series(RandFor.feature_importances_, predictors).sort_values(ascending=False)
coef4.plot(kind='bar', title='Feature Importances')
```

```
Model Results
RMSE : 2960
CV Score : Median - 2962 | Std - 16.62 | Min - 2934 | Max - 2995
```



```
<matplotlib.axes._subplots.AxesSubplot at 0x1a17da5cc0>
```

Figure 53 Random Forest Model [source: own]

Random Forest is also performing very well, with a low RMSE and also close to the median. However, when compared to the Decision Tree, it looks that is the one to perform the best. As the goal of this practical analysis was to find the best model for customer behavior prediction, Decision Tree seems to be the right one. In case of performing and running a complete Machine Learning problem, the next steps would be Hyperparameter Tuning and Ensembling.

6. Conclusions

The reason of choosing this topic for this diploma thesis is behind the personal working experience in an e-commerce company and dealing on a daily basis with customer data. This data included both demographic and behavioral data. While having access and visibility to an immense amount of data, there was a strong personal interest to develop deeper analysis to make data more useful and powerful. Therefore, to gain more insight which could be used from different teams across the company.

As mentioned at the beginning of this thesis, the main goal of it is to develop an in-depth analysis of customer purchase behavior and to determine which of the statistical modeling used is the best one for predicting purchase as target variable. To achieve such goal, partial ones were identified which come in hand as complementary.

The partial goals were covered both in the theoretical part and in the practical part. The literature review starts with a generic overview of Big Data and the common challenges of it. The reason why this is covered is because the dataset used for the analysis is considered to be Big Data, therefore it includes all the challenges and potentials of it. Characteristics such as Volume, Velocity and Variety are covered and pointed as three of the five challenges when dealing with Big Data. One of the main challenges especially when analyzing the data is its Complexity. In the dynamic context of Big Data, the discussion Quality vs Quantity is also something that brings new demands.

In addition to the common challenges, there are analytical challenges as well, like Scaling. One of the primary difficulties when starting with data analysis is getting the right picture of it and the right understanding. Following this, the right interpretation of the data is also part of the statistical challenges which is seen in the practical part during the EDA process.

The data mining techniques used to analyze the dataset are described as part of the literature review, to give an overview on how each of these techniques works and on what type of data. The four models used are: Linear Regression, Ridge Regression, Decision Tree and Random Forest. Other techniques are also mentioned in the theoretical part in order to cover a full picture of how predictive and descriptive modelling work.

In order to best enhance the performance of these techniques, the metric chosen is RMSE, which represents the standard deviation of the differences between the predicted values (theoretical values) and the observed ones.

The dataset includes customer purchase information, so there is an initial theoretical background covered. It includes best practices on how to build the complete customer profile by including demographic and behavioral information and then the techniques on how to predict the probability of customer purchase and on how to best create personalized recommendation for each customer.

The practical part starts with framing the problem, and raising assumptions, which help answer questions and gain insights from the data. The challenges pointed out in the literature review, are captured in the dataset such as missing values, irrelevant variables, data format and many more. Some of these are solved and for some of the others there are recommendations given.

The analysis involved the univariate analysis, where each of the variables was explored independently and bivariate analysis where the target is placed in correlation with the regressors. In this way, some of the questions raised in the assumptions can be answered.

As mentioned during the whole thesis, the main goal of this work, is to find the challenges and opportunities when analyzing customer purchase data and to define which of the techniques selected would be the best one to predict whether a customer would buy from a certain product category in the future. This is done as a Machine Learning problem where after the data transformation, each of the model is trained on the train set and for each of them, a result of the RMSE is produced. After evaluation and comparison, the Decision Tree model is found to be the best one when dealing with customer purchase data. Since the target variable in a decision tree model can be both categorical and continuous, in this case, the purchase amount variable can be predicted in both format: either categorical or continuous.

7. References

Ömer Artun Phd, Dominique Levin. 2015. *Predictive Marketing Easy Ways Every Marketer Can Use Customer Analytics and Big Data.* New Jersey : John Wiley & Sons, Inc., Hoboken, 2015.

Analytics in a Big Data World. **Baesens, Bart. 2014.** 2014.

Big Data for Dummies. **Judith S. Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman. 2013.** New Yersey : s.n., 2013.

Big Data: Issues and Challenges Moving Forward. **Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money, George Washington. 2013.** 2013, p. 998.

Business Statistics. **Kazmier, Leonard. 2004.** 2004, p. 268.

—. **Kazmier, Leonard. 2004.** 2004.

David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied Logistic Regression.* s.l. : John Wiley & Sons, Incorporated, 2013.

Dolliver, Mark. 2009. Comparing Single and Married Women as Consumers. *Adweek.* [Online] June 1, 2009. https://www.adweek.com/brand-marketing/comparing-single-and-married-women-consumers-99425/.

Géron, Aurélien. 2017. *Hands On Machine Learning with Scikit Learn & TensorFlow.* 2017.

—. 2017. *Hands-On Machine Learning with Scikit-Learn & TensorFlow.* 2017.

Gooddata Help. *Normality Testing - Skewness and Kurtosis.* [Online] [Cited: January 28, 2019.] https://help.gooddata.com/display/doc/Normality+Testing+-+Skewness+and+Kurtosis.

Hands-On Machine Learning with Scikit-Learn & TensorFlow. **Géron, Aurélien. 2017.** 2017, p. 26.

Harvey, Cynthia. 2017. Big Data Challenges. *Datamation.* [Online] 2017. https://www.datamation.com/big-data/big-data-challenges.html.

Hill, Kashmir. 2012. How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did. *Target.* [Online] February 2012. https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#4daec64c6668.

International Encyclopedia of Education (Third Edition). **Sinharay, S. 2010.** 2010.

Ismail, Nick. 2017. *Information Age.* [Online] November 8, 2017. https://www.information-age.com/real-opportunities-big-data-digital-world-123469428/.

Khade, Anindita A. 2016. *Performing Customer Behavior Analysis using Big Data Analytics.* s.l. : Elsevier B.V., 2016.

Optimove. 2018. Customer Behavior Modeling. *Optimove.com.* [Online] October 2018. https://www.optimove.com/learning-center/customer-behavior-modeling.

Ridge Regression. *NCSS Statistical Software.* [Online] https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf.

SAS Insights. Big Data What it is and why it matters. *www.sas.com.* [Online] https://www.sas.com/en_us/insights/big-data/what-is-big-data.html.

TechAmerica Foundation's Federal Big Data Commission. *DEMYSTIFYING BIG DATA.*

The SAGE Encyclopedia of Communication Research Methods. **Allen, Mike. 2017.** s.l. : SAGE, 2017.

UN Global Pulse. 2012. *Big Data for Development: Challenges & Opportunities.* 2012. pp. 24-33.

Unruh, Heidi. 2018. UNDERSTANDING WHY MICRO-SEGMENTATION IS CRITICAL TO DIGITAL EXPERIENCES. *Chief Marketer.* [Online] June 2018. https://www.chiefmarketer.com/understanding-micro-segmentation-critical-digital-experiences/.