ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

PROVOZNĚ EKONOMICKÁ FAKULTA

KATEDRA STATISTIKY

Aplikace credit scoringového modelu v bankovní praxi

Bakalářská práce

Vedoucí práce:
Ing. Tomáš Hlavsa, Ph.D.

Vypracovala:
Helena Dobešová

Abstrakt

Bakalářská práce se zabývá tvorbou credit scoringového modelu, který je aplikovatelný v praxi. Na základě dataminingového procesu byly vytvořeny tři dílčí modely na třech datových sadách, které jsou postaveny na datech v minulosti a předpovídají chování klientů v budoucnu. Jednalo se o model aplikační, kreditní a behaviorální. Dále byly nalezeny signifikantní proměnné, které mají na rozhodnutí o poskytnutí úvěru největší vliv. V metodické části je popsán kompletní postup výstavby prediktivního modelu a celé analýzy. Následně je objasněna nejdůležitější terminologie, kde je kladen důraz na vysvětlení dat, která banky sbírají, a rozdělení klientů. Poté jsou detailně popsány jednotlivé modely, jimiž musí klient v procesu schvalování projít. Pro účely analýzy klientských dat byly využity statistické metody explorační analýzy, prediktivní analýzy a regresní analýzy. Na základě těchto výsledků byly vytvořeny tři scorekarty a tři rovnice logistické regrese. Celý proces byl validován na testovacích datech. Nakonec byly modely demonstrovány na konkrétních klientech a tím potvrzeno jejich praktické využití. Na základě zjištěných informací a výsledků testů byla navržena možná praktická zlepšení celého úvěrového procesu z pohledu banky i klienta.

Klíčová slova

Credit scoring, úvěr, žadatel, prediktivní model, logistická regrese, rozhodovací strom, Gini index, Weight of evidence, rating, behaviorální skóre, aplikační skóre, credit bureau

Cíl práce

Cílem této práce je vytvoření credit scoringového modelu. Dílčím cílem bude vytvoření tří modelů na třech datových sadách, které budou aplikovatelné v praxi a budou mít co možná nejlepší vypovídající charakter. Simultánně se bude práce zabývat slabinami celého úvěrového procesu a modelu, se kterými je možné se potkat v praxi. Dále budou navržena možná vylepšení celého úvěrového procesu. V neposlední řadě se bude práce zabývat nalezením vhodného nastavení cut-off hladiny, aby banka co nejvíce vytěžila tržní potenciál za současné

minimalizace ztrát. V závěru budou představeny proměnné, které by mohly mít největší vliv na predikci modelu. Nedojde však ke kompletnímu odhalení všech proměnných, které banky sbírají, aby nedošlo k vyzrazení know-how.

Dalším cílem je seznámení čtenářů s hloubkou celého procesu a zorientování v problematice. Celý proces bude demonstrován na třech náhodně vybraných klientech s kompletním výsledkem prediktivní analýzy. Ve výsledcích bude demonstrován ideální klient pro banku z každého datasetu.

Metodika

Celá práce se dělí na dvě hlavní části. První část se zabývá teoretickými východisky. Druhá část je zaměřena na vlastní práci, ve které je vyhotoven credit scoringový model a zpracovány tři datové soubory a vyhotoveny tři modely a tři scorekarty. První datový vzorek byl použit na vyhotovení aplikačního modelu, druhý na kreditní model a třetí na model behaviorální.

Pro účely analýzy klientských dat bylo nejdůležitější pochopit celý datový soubor. Dále bylo nutné se vypořádat s chybějícími daty a odlehlými hodnotami, kde bylo využito grafické znázornění pomocí krabicových grafů (Box plotů). Následovalo rozdělení datového souboru na dva vzorky. Jeden testovací, druhý trénovací. Na trénovacím vzorku byl sestaven model a na testovacím na základě metody Cross-Validace potvrzena prediktabilita celého modelu.

Následně bylo zkoumáno, jak se jednotlivé proměnné chovají (Intuitive Behaviour) pomocí metody Weight of Evidence. Na základě této metody byla každá proměnná rozdělena na jednotlivé kategorie (biny). Dále proměnné vstupovaly do single faktor analýzy kde byla zkoumán jejich vypovídající hodnota (Information Value) a jejich síla pomocí indexů Gini, Kolmogorov-Smirnov Index, Somers'D index a ROC křivky.

Před finálním vstupem do modelu byly spojité proměnné testovány pro případný výskyt multikolinearity pomocí Pearsonových korelačních koeficientů uspořádaných do korelační matice. V případě kategoriálních proměnných byly pro identifikaci multikolinearity využity Pearsonovy kontingenční koeficienty. Následně rozbinované proměnné vstupovaly do multivariantní analýzy, kde byly pomocí logistické regrese vybrány signifikantní proměnné. Na základě těchto přístupů byly vybrány proměnné, které dosahovaly hodnoty p nižší než 0,05. Výsledkem celého procesu byla scorekarta, ve které jsou uvedeny jednotlivé koeficienty logistické regrese pro statisticky významné proměnné.

Teoretická část

Teoretická část ze zabývá vymezením veškeré nutné terminologie, která byla zpracována na základě odborné literatury a podrobného studia dostupných pramenů. Důraz je kladen především na vysvětlení dělení bankovních dat a rozdělení klientů. Dále na objasnění úvěrů, pro které byl model sestaven a na vysvětlení pojmu Credit scoring.

Praktická část

V praktické části byl sestaven credit scoringový model. Postup prediktivní analýzy byl detailně demonstrován na aplikačních datech a následně aplikován na datech kreditních a behaviorálních. Testována byla především nejlepší kombinace modelu, která se skládala z proměnných rozbinovaných na základě metody Weight of Evidence a logistická regrese metodou Forward Stepwise nebo Backward Stepwise. Dále bylo testováno praktické využití indexů Gini, Kolmogorov-Smirnov a Somers'D. Modely byly validovány na základě Cross-Validace, kde bylo nutné nastavit správnou hladinu cut-off tak, aby odpovídala strategii banky a zároveň byl maximalizován zisk za současného řízení rizika a ztrát. Na konci došlo k potvrzení praktického využití na základě náhodně vybraného klienta, kde se potvrdila správná prediktabilita všech modelů

Závěr

Cílem této práce bylo sestavit na základě statistické analýzy credit scoringový model s nejlepšími prediktivními vlastnostmi. Dílčím cílem bylo nalezení nejlepšího modelu a nejlepšího přístupu k analýze. Na základě rozsáhlého testování byly sestaveny modely a vybrán takový, který nejlépe odhadoval chování klientů v budoucnu.

Z uvedených výsledků se u všech tří modelů osvědčil přístup v kombinaci s Long listem proměnných, WOE kategoriálních proměnných a logistické regrese pomocí metody Backward Stepwise.

Celkem vyšlo 13 signifikantních proměnných na aplikačním datasetu, u kreditního bylo nalezeno 12 signifikantních proměnných a u modelu behaviorálního 15 signifikantních proměnných.

Na základě ROC křivek byl prokázán behaviorální model jako nejsilnější.

Z toho plynulo, že pokud by měla banka porovnat informace, které získá z behaviorálních dat a aplikačních dat, vyšly by jednoznačně prediktivnější a silnější data behaviorální. Proto, pokud nebankovní klient požádá o úvěr u jiné banky, než je jeho domovská, předkládá bance výpisy z účtu. Banka by s nimi měla umět následně pracovat. Možností by bylo elektronické čtení výpisů z účtu z jiné banky, který klient donese jako dokumentaci k úvěru. Pokud by se bance podařilo takto nasimulovat data, získala by lepší výsledky predikce a měla by možnost se lépe u klientů rozhodnout.

Dalším důvodem pro sběr dat touto formou může být zatajení některých informací ze strany klienta. Pokud by například hrál hazardní hry nebo sázel, banky hodnotí takový fakt velice negativně. Často z tohoto důvodu klienta rovnou zamítnou. Stejné platí pro fakt, že klient může zatajit počet dětí, které má, popř. zatajit úplně, že děti má. Každá osoba bez příjmu navyšuje existenční výdaje a snižuje potencionální výši limitu úvěru. Z účtu se dá ale vyčíst, jestli platí výdaje za stravné, školu nebo například alimenty. Pokud by klient informace zatajil, úvěr banka zamítne pro možný pokus o úvěrový podvod.

Další potenciální hrozba pro klienty je počet otevřených žádostí. Často, když se klient rozhodne zažádat o úvěr, obejde hned několik bank, kde podepíše žádost o úvěr.

Z výsledků je patrné, že proměnné *NumOfRejProd_off_us* a *NumOfProd_2y* (*počet zamítnutých žádostí a počet žádostí celkem*), jsou signifikantní a rozhodují finální stanovisko banky. Proto si klient s každou další žádostí o úvěr škodí a snižuje tím svůj rating. Z toho důvodu není dobré obcházení více bank a otvírání velkého množství žádostí. Stejné platí pro makléře, kteří s klienty banky obchází, čímž klienta poškozují.

V závěru práce byla v rámci diskuze navržena sada zlepšení celého procesu. V neposlední řadě mají doporučení i lidský rozměr. Neposkytnutím úvěru nebonitním klientům banka předchází možnému vzniku krizových životních situací typu exekucí. Současný trend regulací České národní banky se také snaží těmto krizovým situacím předejít.

Seznam zdrojů a použité literatury:

Seznam použité literatury:

1) SIDDIQI, Naeem. *Intelligent Credi Scoring.* New Jersey: John & Sons, Inc., 2017. ISBN 978-1-119-272915-0.

2) ABBOTT, Dean. *Applied Predictive Analytics.* USA: John Wiley & Sons, Inc., 2014. ISBN 978-1-118-72796-6

3) TUFFÉRY, Stéphane. *Data Mining and Statistics for Decision Making*. UK, West Sussex: Wiley, 2011. ISBN 978-0-470-68829-8.

4) CIPRA, Tomáš. *Riziko ve financích a pojišťovnictví: Basel III a Solvency II*. Praha: Ekopress, s. r. o., 2015. ISBN 978-80-87865-24-8.

5) CIPRA, Tomáš. *Praktický průvodce finanční a pojistnou matematikou*. Praha: Ekopress, s. r. o., 2015. ISBN 978-80-87865-18-7

6) FINLAY, Steven. *Credit Scoring, Response Modelling and Insurance Rating.* USA: Palgrave Macmillan, 2010. ISBN 978-0-230-57704-6