bank-data-analysis-answers

September 23, 2022

1 Bank Data Analysis

We have 5000 rows (samples) and 13 columns (features). One row should correspond to one applicant. The features: - CODE_ZIP – zip code of location of applicant, 1430 unique zip codes

- AMT_NET_INCOME - average monthly income in crowns, mean is 27041.529 and median is 21000, this should most likely represent income after taxes
- AMT_REQUESTED_TICKET - average monthly requested ticket
- TEXT_BANK - bank of the appliccant, we have 14 banks with most in Ceska sporitelna
- NUM_AGE - numeric age from 18 to 84
- TEXT_GENDER - gender in format "Muž" "Žena" with more men
- NFLAG_MOBILEDEVICE - flag if applicant uses mobile device, more applicants use it (approx 2:1 for use:dont use)
- CODE_IP_1 - no idea what this means
- NUM_LEVEN_EMAIL - no idea
- NFLAG_EMAIL_NUMERAL - whether applicant uses email
- CNT_REJECTED - I guess this should be whether last one was rejected?
- NUM_DAYS_CREDIT_HISTORY - number of days of credit history
- TARGET - whether loan wasnt paid on time and properly

[Q1] What's the highest age of applicants? (Jaký je nejvyšší věk žadatele?)

[A1] Highest age is 84 years. We just take the maximum value of NUM_AGE column.

[Q2] What is the mode age of applicant? (Jaký je modus věku žadatele?) Mode age of applicant is 20 years. Here we take the age that occurs most often.

[Q3] What is the average income of applicant and how does that differ fro the average in Czech republic? (Jaký je prúměrný příjem žadatele a jak se liší od prúměru v ČR?)

[A3] Average income of applicant is 27041.529. This should represent income that arrives to applicants bank account. We could calculate is as sum of average income for each applicant divided by number of applicants. 40086 is the average monthly GROSS income in Czech republic (https://www.czso.cz/csu/czso/cri/prumerne-mzdy-2-ctvrtleti-2022). I wasnt able to find information about salary after taxes (čistá mzda), thus I use online calculator

(https://www.vypocet.cz/cista-mzda) to calculate it from the found gross income (we ignore kids/students etc..). The result is: 32231. Thus the average monthly income in Czech republic currently is higher than the one in the dataset. However this can be caused by me using the most recent statistics while the dataset contains bit more historic data.

[Q4] In what age groups are more men than women? (Ve kterých věkových skupinách (18-28, 29-38, 39-48, ..) je více mužú než žen?)

[A4] In age groups 18-28, 29-38, and 39-48. I calculated this by creating age bins and then contingency table to count all possible combinations of age bins with gender.

[Q5] Ve které bance je největší target rate (tedy poměr počtu TARGET=1 na celku)?

[A5] In bank BNP Paribas Personal Finance SA, odštěpný závod. We create contingency table, normalize it and take the one with highest ratio.

[Q6] Try to find ideal age categories to analyze target rate. (Pokuste se vytvořit vhodné věkové kategorie pro zkoumání target rate.)

[A6] We used the optimal binning library which uses mathematical programming to find the optimal bins regarding the target variable (see https://arxiv.org/pdf/2001.08025.pdf and https://github.com/guillermo-navas-palencia/optbinning). The found bins are 0-19, 20, 21-23, 23-26, 26-30, 31-36, 37-41, 41-51, 52+. The highest target rate is for 0-19 and lowest for 37-41.

[Q7] How does target rate differ for different banks according to age? (Jak se liší výsledky target rate podle jednotlivých bank v závislosti na věku?)

[A7] We can leverage our binned age and bank to create contingency table and then extract the target rate (ratio of True target to False+True). Further we can create barplots here (where y is target rate and x/hue are age bin and bank). Then we can inspect the numbers and see differences: - for example air bank, and fio have the highest target rate for people up to 19 years old (including) - obchodni banka has highest target rate for 20 years old - raiffeisen for people from 52 years - equa bank for 37-40 years old Further we could focus more on some banks and look at the trends more or even inspect histograms of age per target and bank (two histograms per each bank). We could also leverage two-way ANOVA to test difference between numerical age according to interaction between TARGET and BANK. But I think the plots tell us more about the differences in target rate.

[Q8] Design the best possible rules that we can use for variables to choose 4-6% records for them to have the best possible target rate. (Navrhněte co nejlepší pravidla, jak pomocí vysvětlujících proměnných ze vzorku vybrat 4-6% záznamů tak, aby ve vybrané sadě byl co největší target rate.)

[A8] Here we experiment with different simple rules and adding them together. For the numerical values I looked at box plots of NUM_DAYS_CREDIT_HISTORY, AMT_REQUESTED_TICKET, AMT_NET_INCOME. I could see most above 10000 requested ticket are approved, and more under 1615 credit history are approved, while there are some outliers that above 80000 income are only rejected. We create three boolean rules out of these values. Then we can also look at other categorical features like binned age, gender, .. Next we just calculate target rates for some promising candidates while moving down to 5% of samples. We have two promisign results: - taking only women are up to 19 years old achieves 21% target rate - takingn people with less credit history than 1615, more income than 10000, age up to 19 years old, mobile device, and CNT_REJECTED we achieve 19.5% target rate

[Q9] (Obohaťte data tak, že přidáte jeden nový atribut odvozený z PSČ, tedy třeba počet obyvatel, okres, kraj nebo cokoliv jiného. Můžete použít jen veřejné zdroje, ty uveďte.)

[A9] We add data from ceska posta (https://www.ceskaposta.cz/ke-stazeni/zakaznicke-vystupy) which should be reliable. However some PSC is missing there (invalid data? new psc?) so we replace it with special UNKNOWN value.

[Q10] Which mathematical model would u recommend to model TARGET variable based on other variables and why would you choose this method? If u want u can create model and comment it. (Jaký matematický model byste doporučili pro modelování cílové proměnné TARGET za základě ostatních vysvětlujících proměnných a proč byste zvolili právě tuto metodu? Pokud máte čas a chuť, tak nějaký model vytvořte a okomentujte jeho vhodnost po použití při schvalování úvěrů)

[A10] We dont have much data available and in this task we should focus on interpretability of our model. Thus I would choose model that can be easily explained to the "customer". My choices would be either logistic regression or nearest neighbours classifier. Obviously we would need to split our data into train and test sets (we should have actually done that right in the beginning before analysis if we know we gonna model). For the features, we would take the ones we already know work (age, bank), the ones which are logicaly needed (income and requested ticket) as our baseline. Then we would consider other features (if they are not correlated like mobile phone with age might be) andd if we can actually legally use the features (can gender and okres be used?). For evaluation metric we would need to take into account the disbalance of accepted and rejected loans by not using accuracy.

1.1 Conclusion (feedback)

Analysis took me 5hours. Issue was that I started the analysis, suddenly I didnt have time.. then I went back to the analysis week later. This meant I had to once again get into the project. The most frustrating was question 8. I had no idea whats the best way to find ideal rules (besides experimenting or doing some bruteforce). Also question 7 was quite hard (I had issue thinking how to decide and answer this).