Czech Bank - Loan Analysis

Moe Khan

March 24, 2021

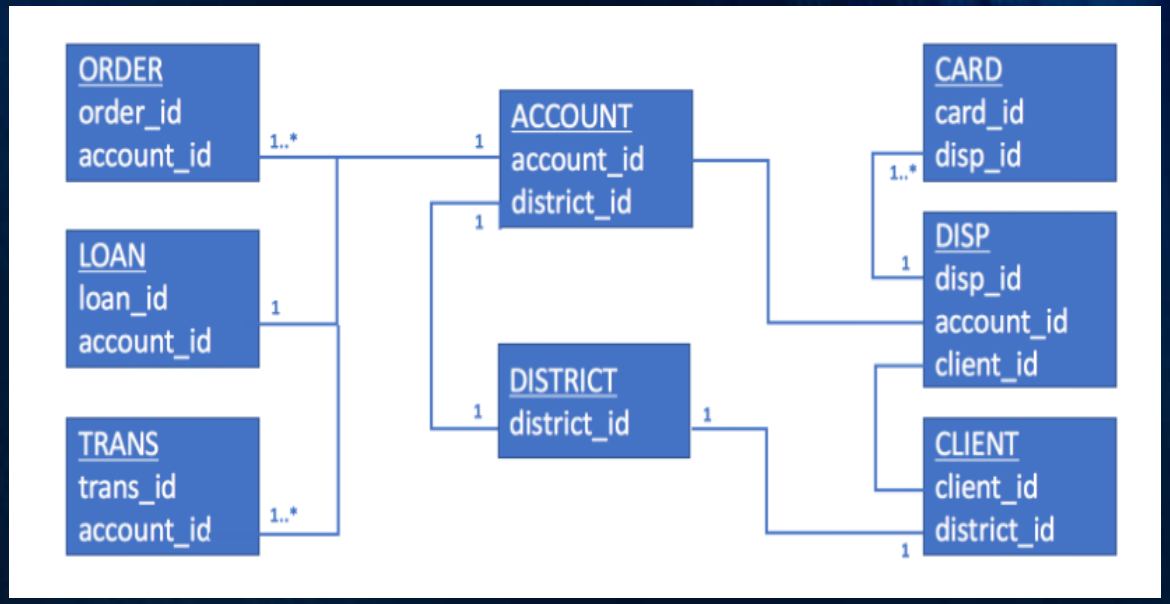
CNB CZECH NATIONAL BANK

Objective:

Predicting loan behavior by analyzing transactional data with classification models.

Research Questions

- Which accounts are likely to default on loan? Why?
- What are the characteristics of a good client?
- What are the top 5 features of a good loan?
- Is there any difference in loan default among branches?



Dataset consists of the following tables:

Account – Each record describes static characterizes of an account (4500 accounts)

Client – Each record describes static characterizes of a client(5369 clients)

Disposition – Describes whether client is owner or user. (5369 dispositions)

Orders – Each record describes payments(6471 orders)

Demographic – Each record describes bank branch in Czech (77 branches)

Card – Each record describes credit card issued to the client (892 cards)

Loans – Describes loan granted to the the account. (682 loans)

Transactions – Each record describes transaction on an account(1056320 transactions)

Data Cleaning process for Machine Learning

Creating Tables in SQL

↓

Connect SQL in Jupyter

↓

Exploratory Data Analysis

↓

Data Cleaning

↓

Feature Engineering

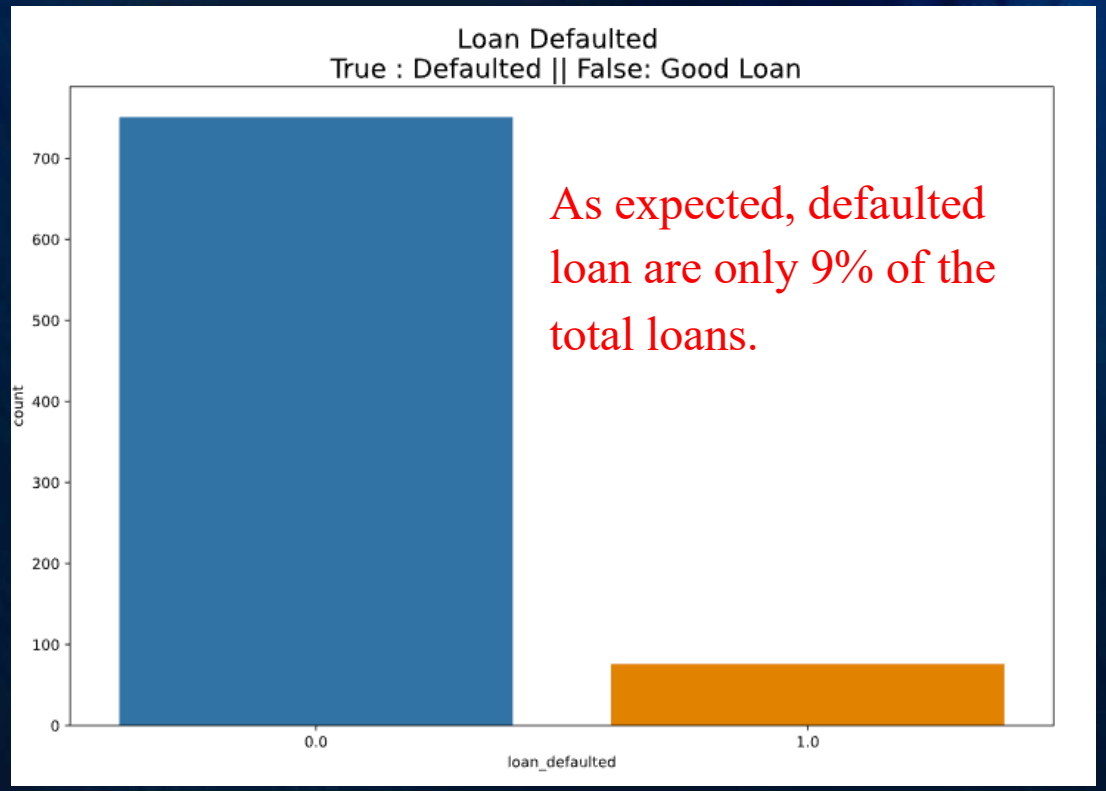
↓

Modelling

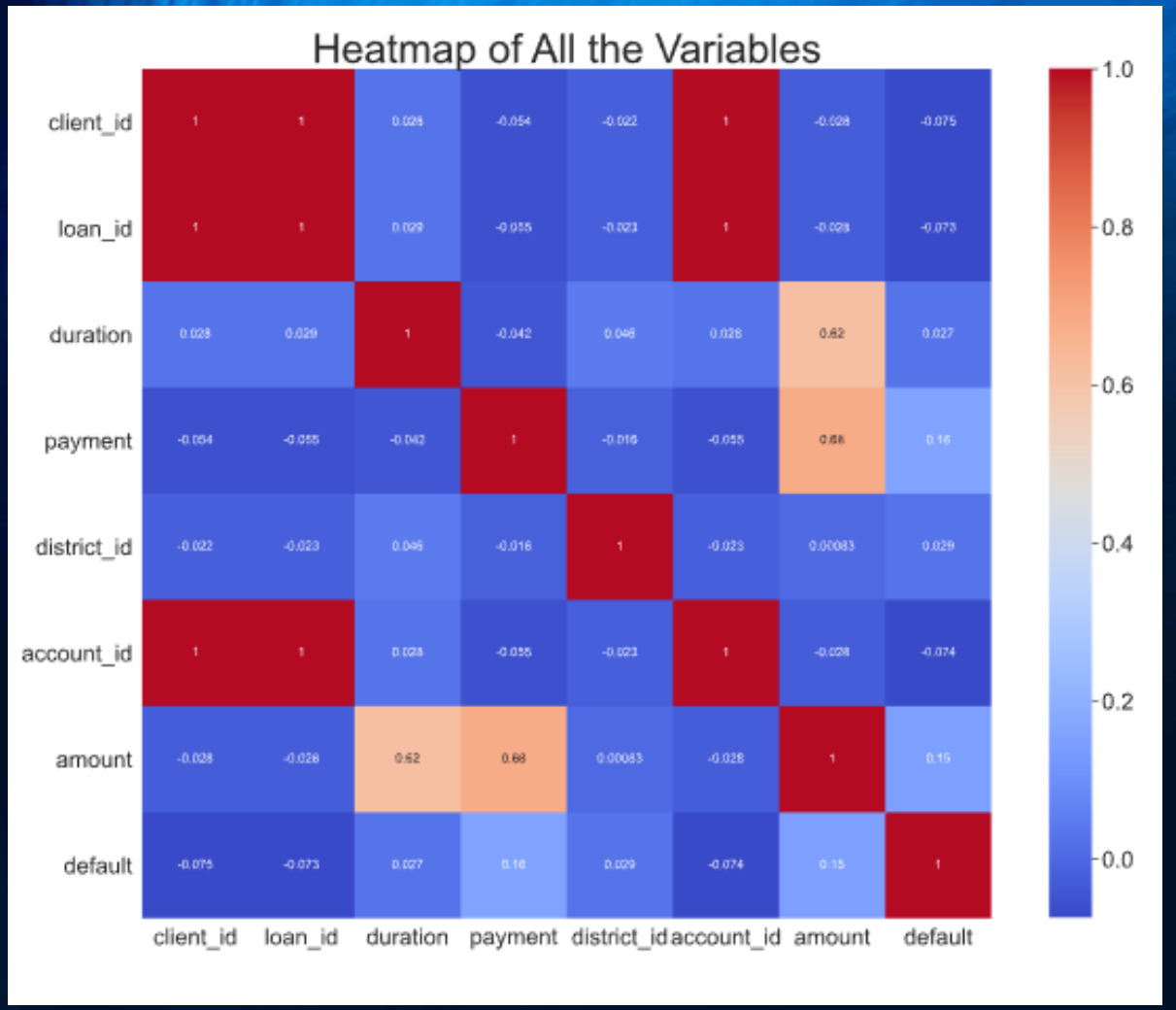
↓

Evaluation

EDA (Extrapolatory Data Analysis)

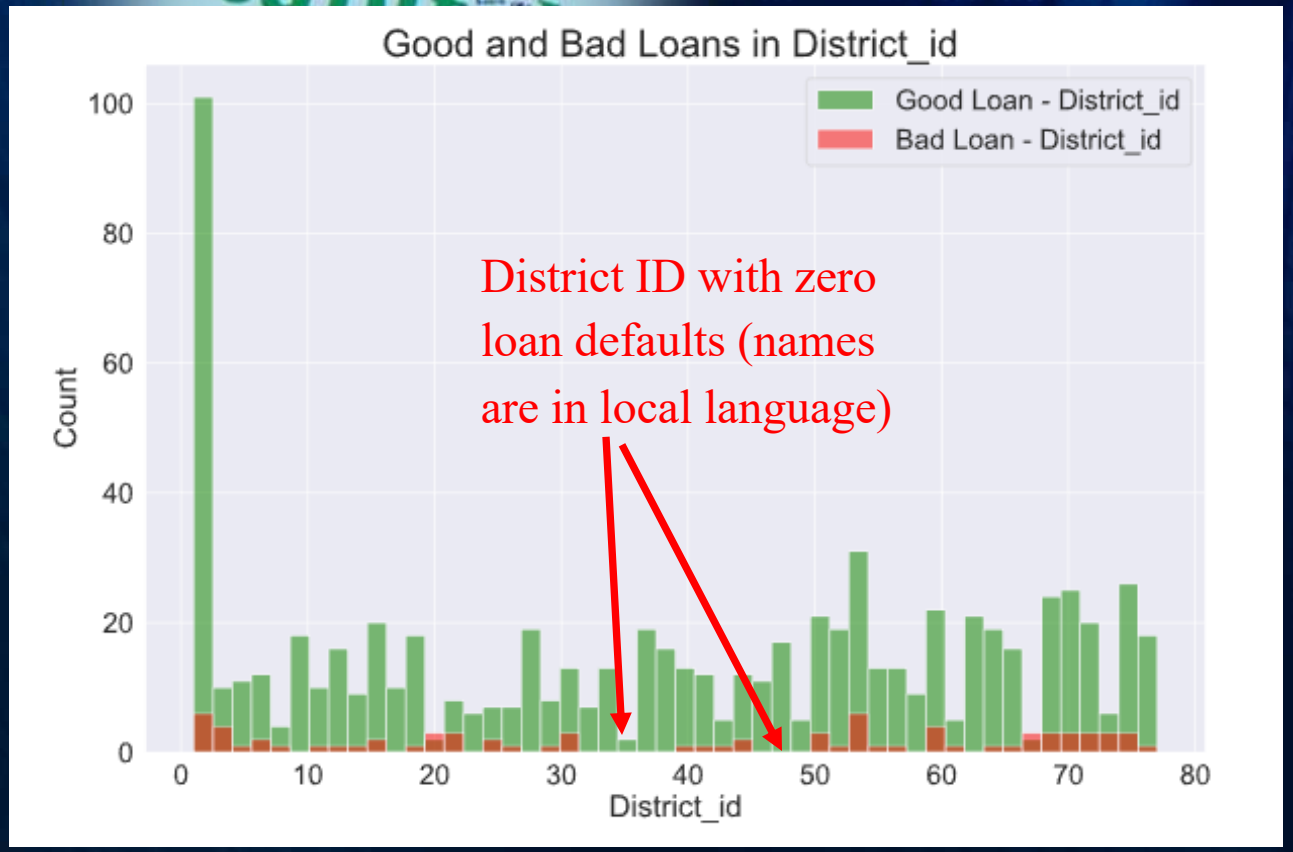


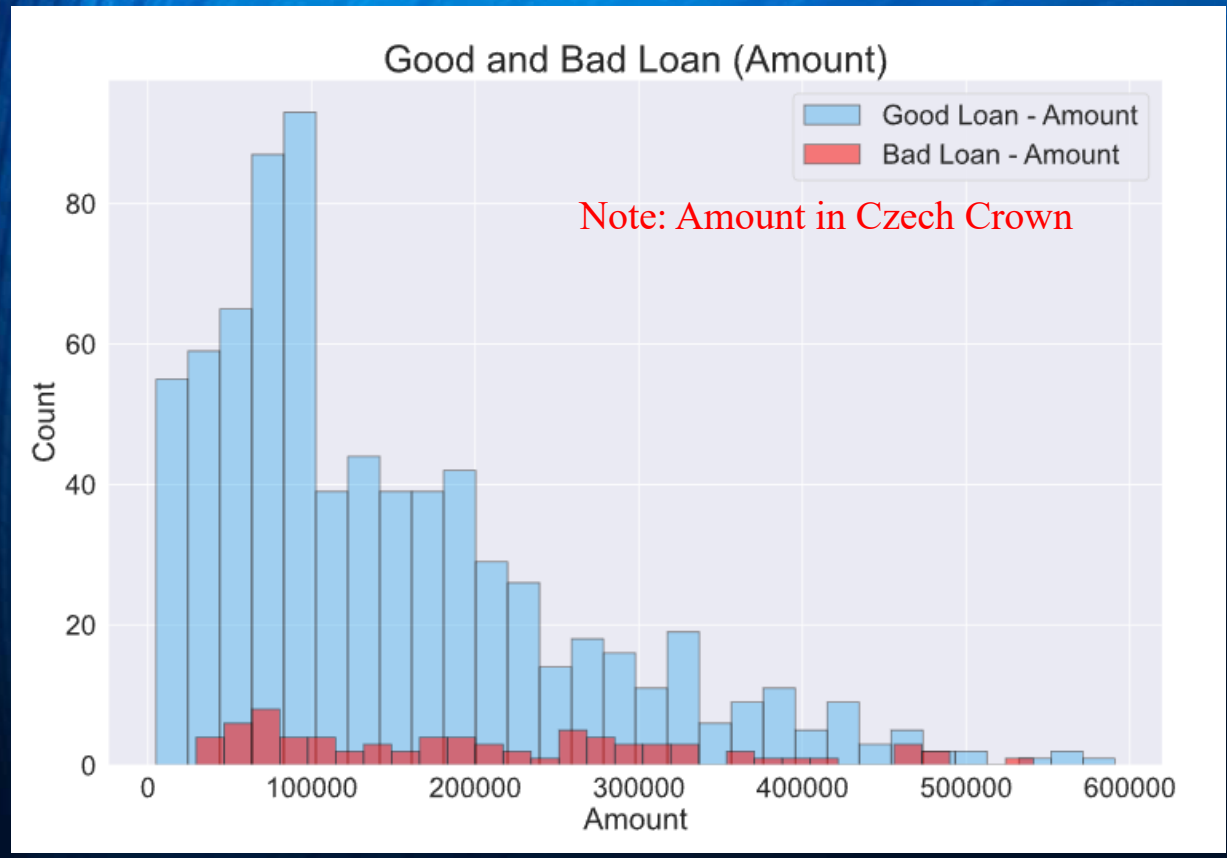
As expected, defaulted loan are only 9% of the total loans.



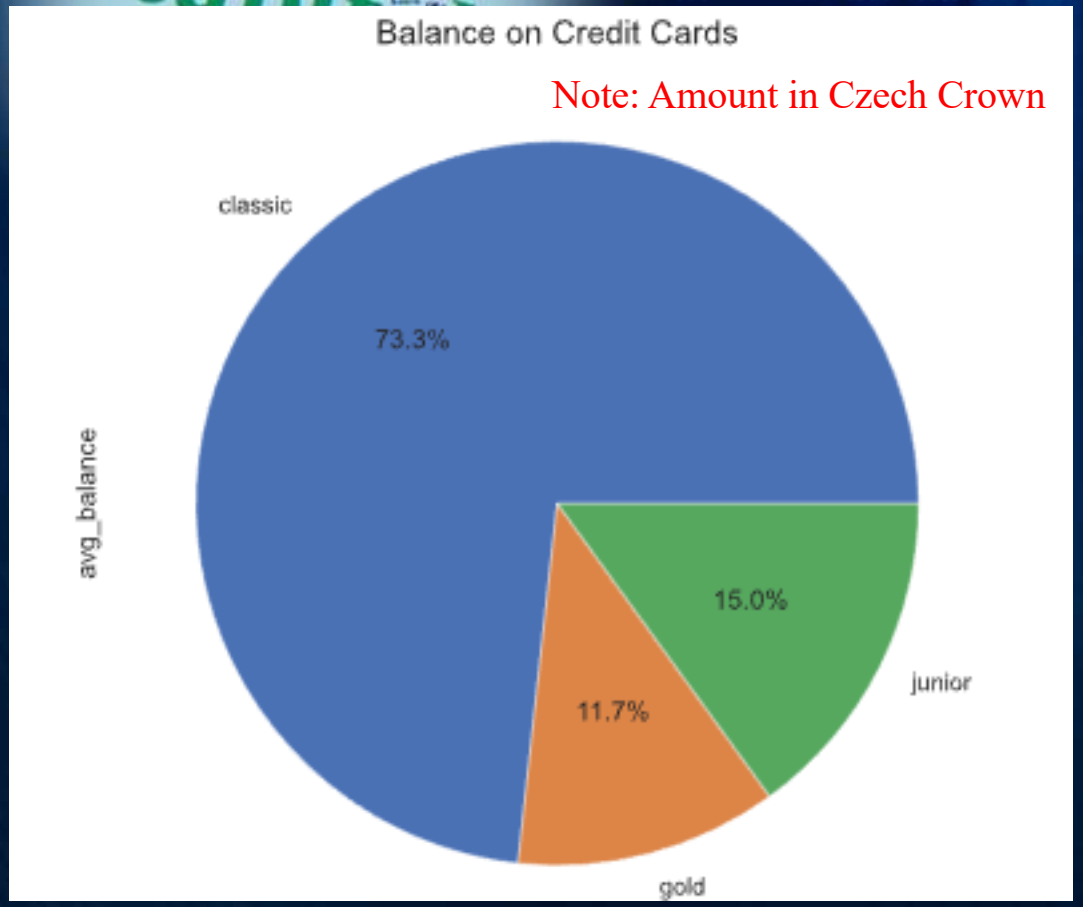
Strong correlation between duration and payment. Only one feature will be considered during feature engineering.

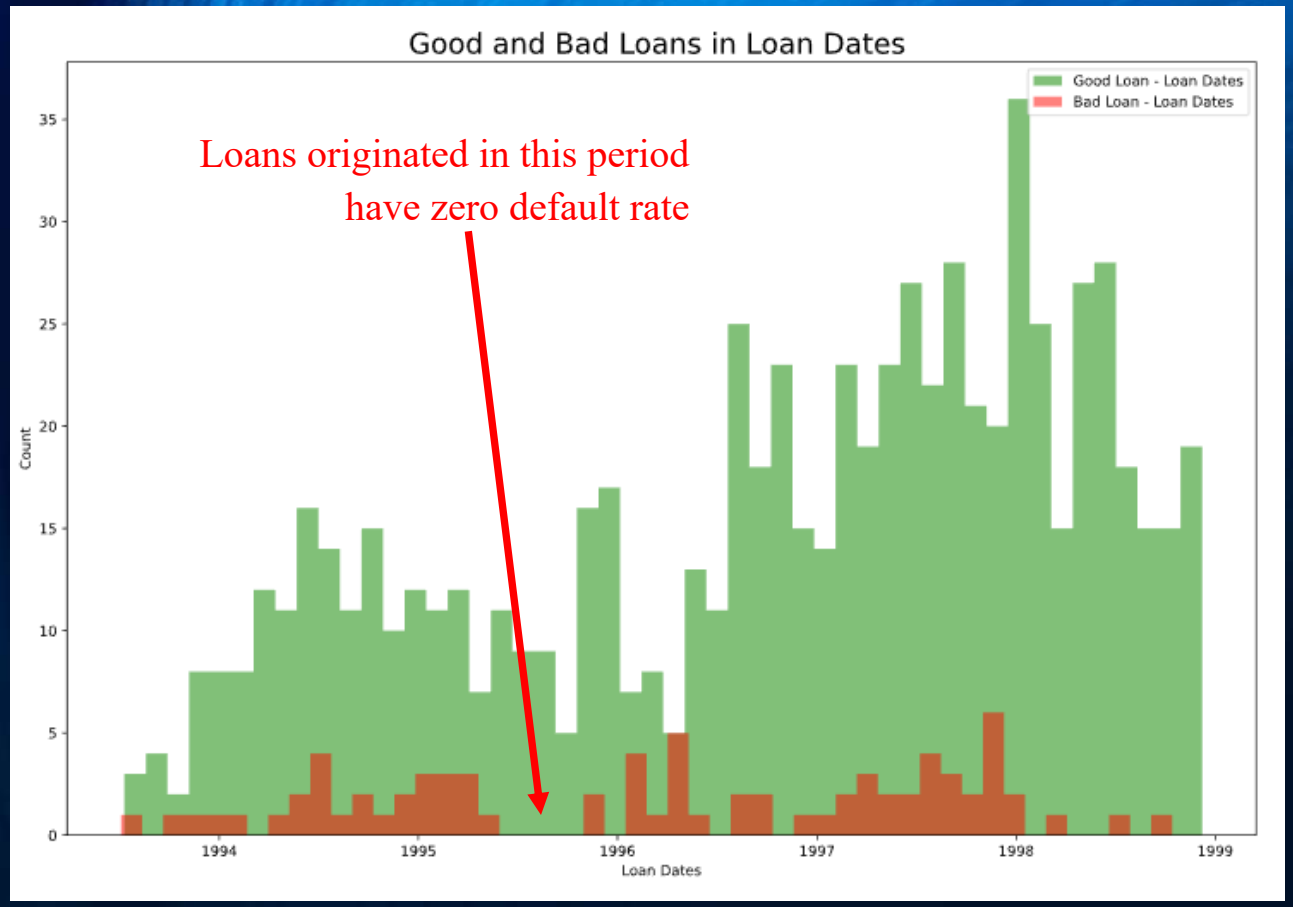
EDA (Extrapolatory Data Analysis)



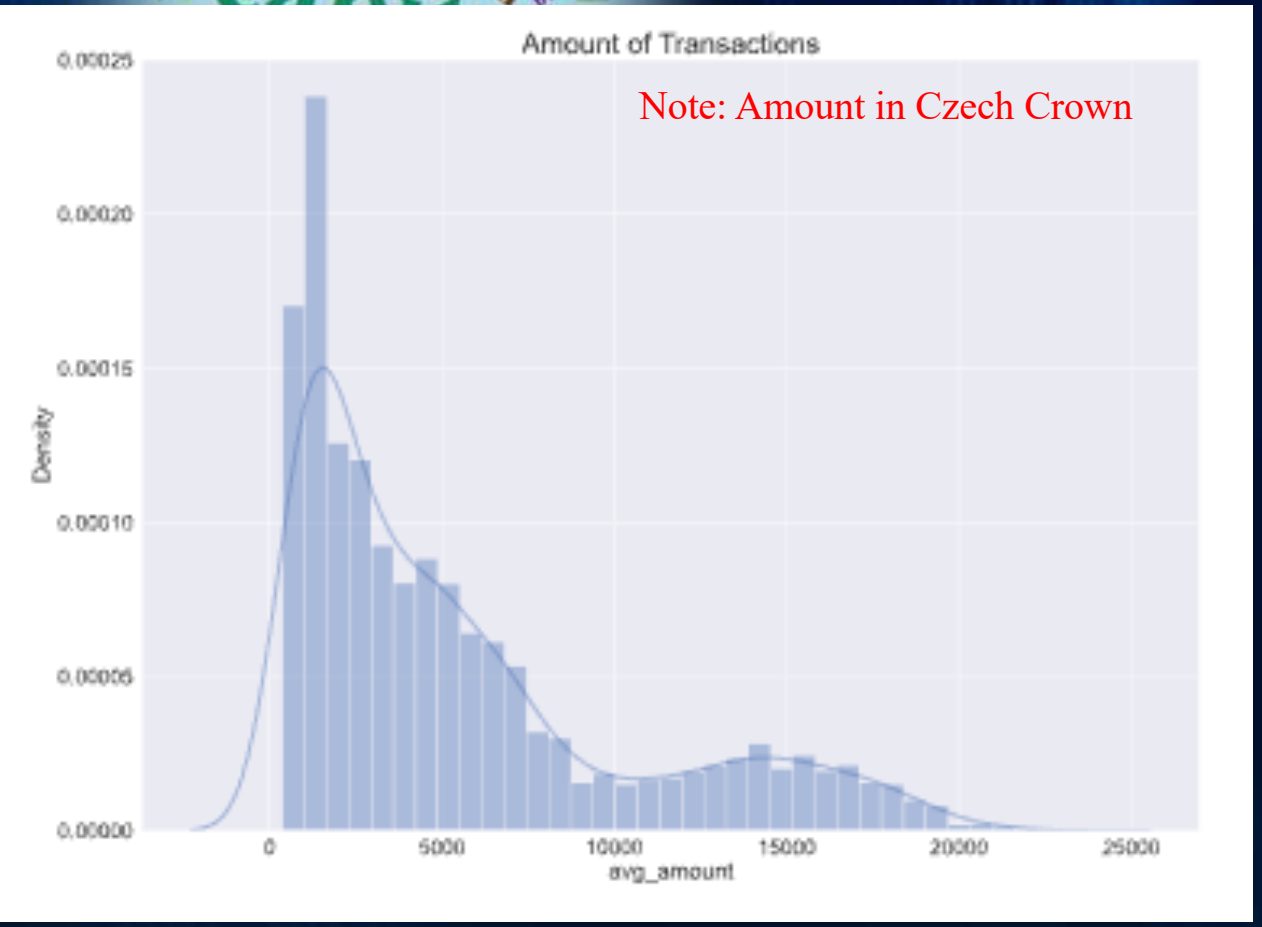


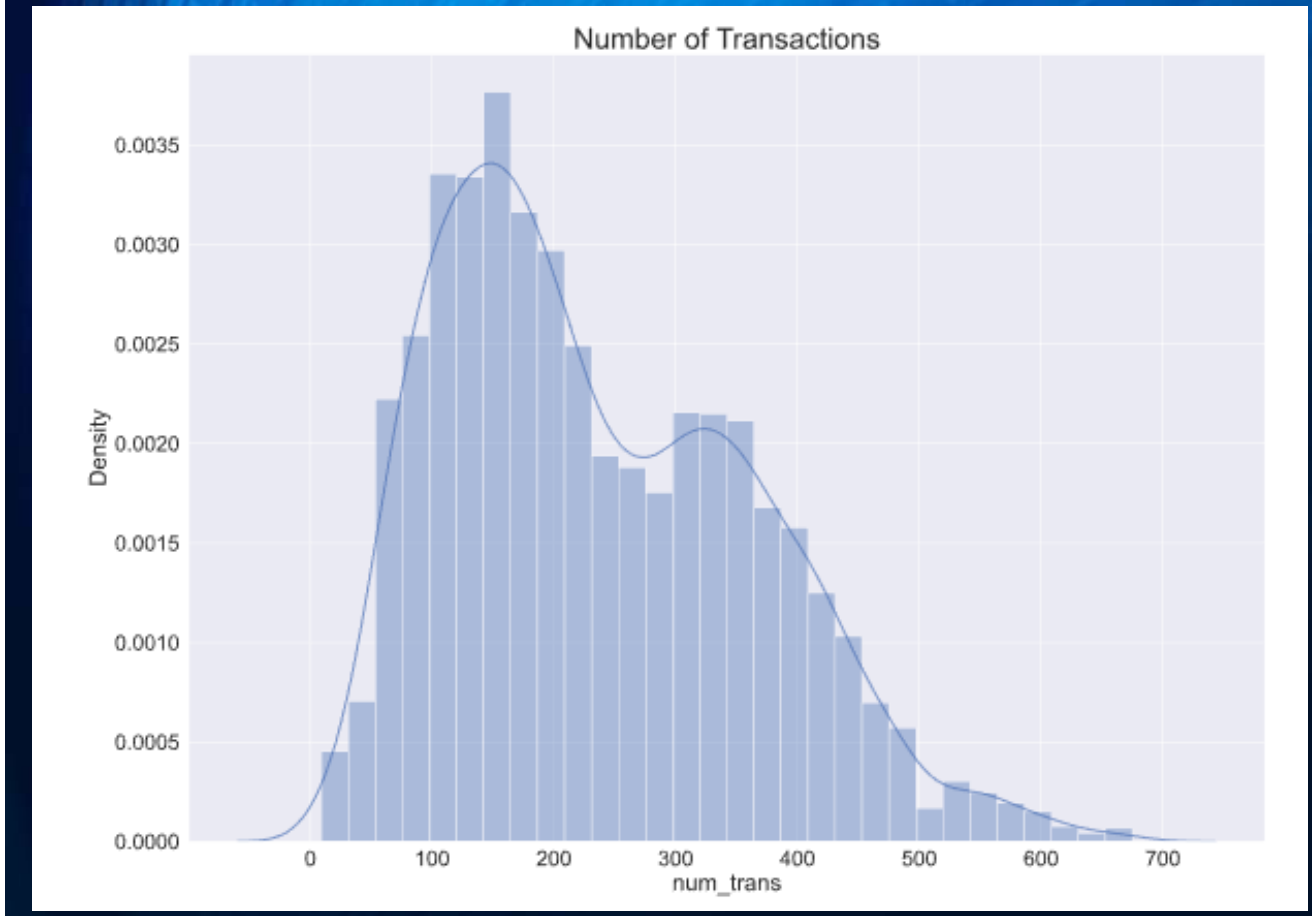
EDA (Extrapolatory Data Analysis)





EDA (Extrapolatory Data Analysis)

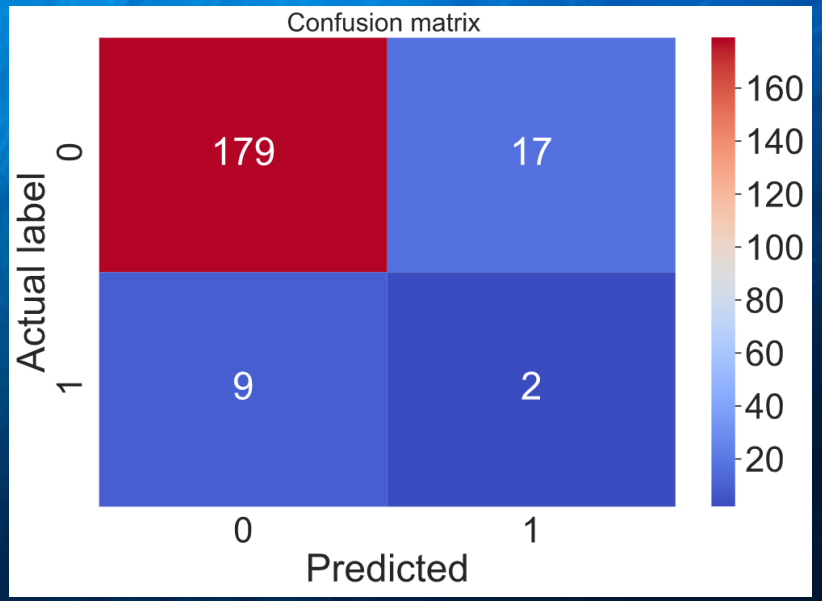


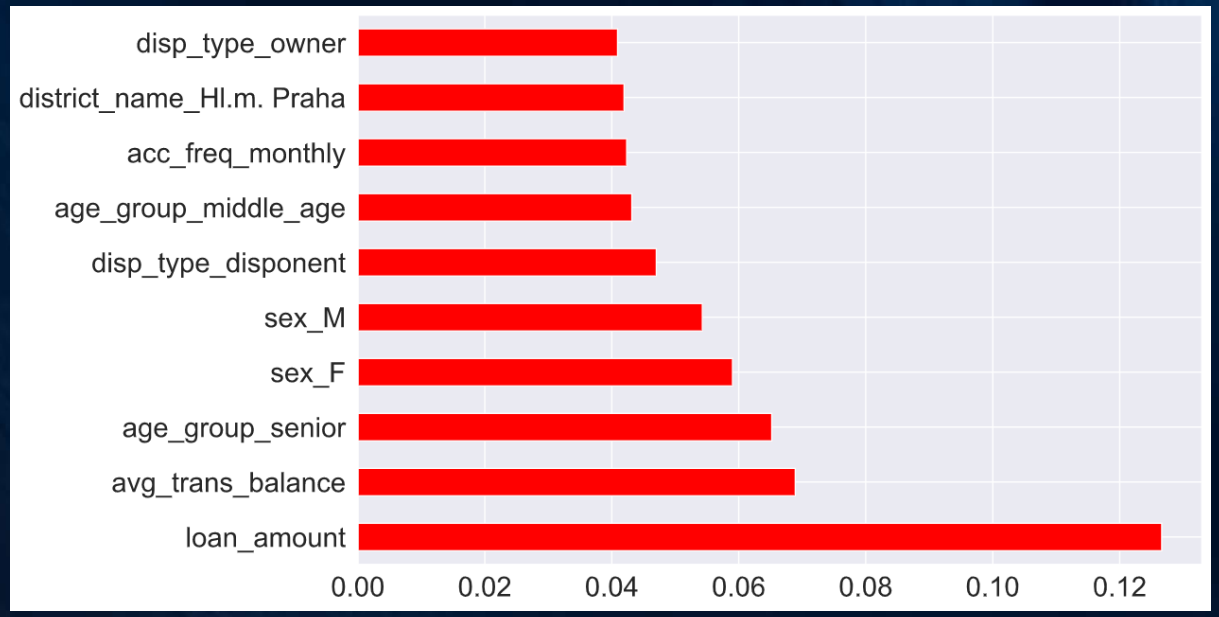


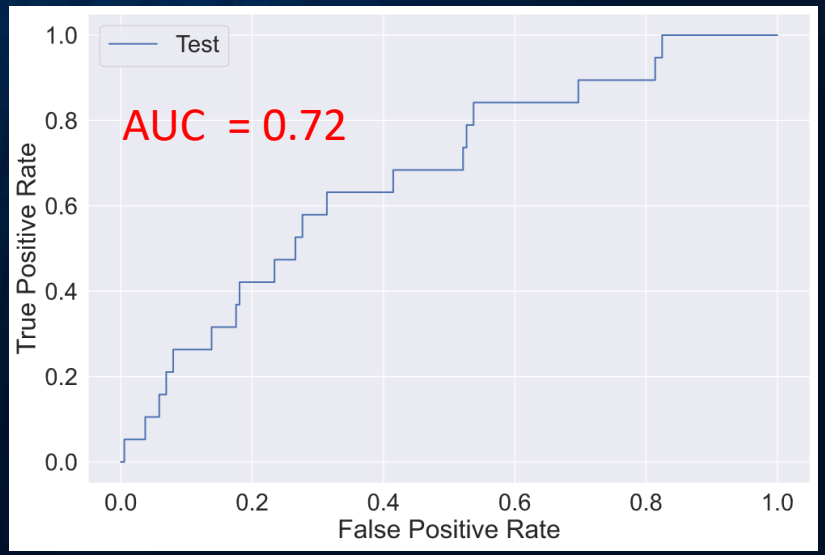
Loan Analysis – Random Forest with SMOTE

Findings:

- Client with the highest loan amount are more likely to default.
- Higher the transaction balance, higher the default rate.
- Seniors are more likely to default than junior or middle age.
- Male and Female clients have the same probability of default.



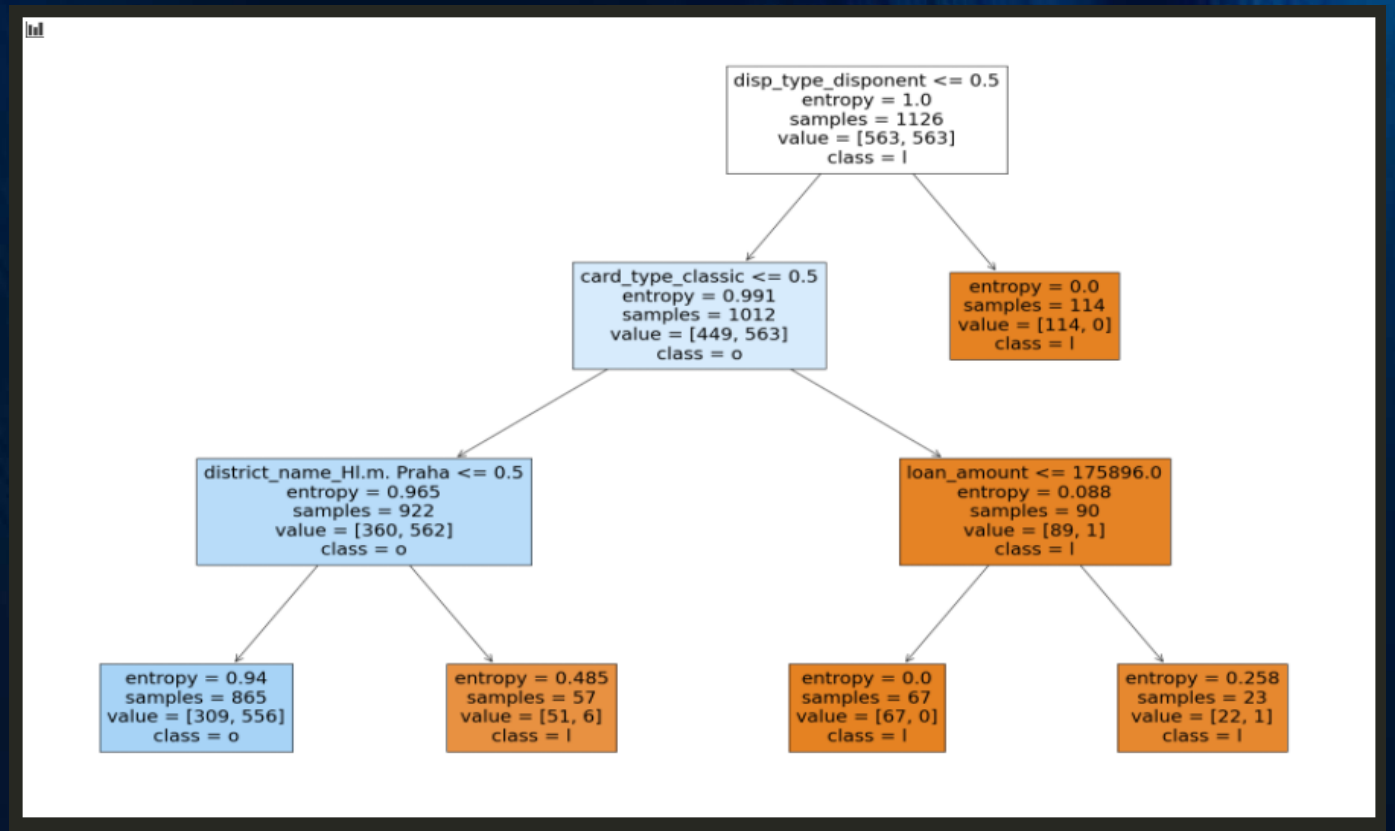


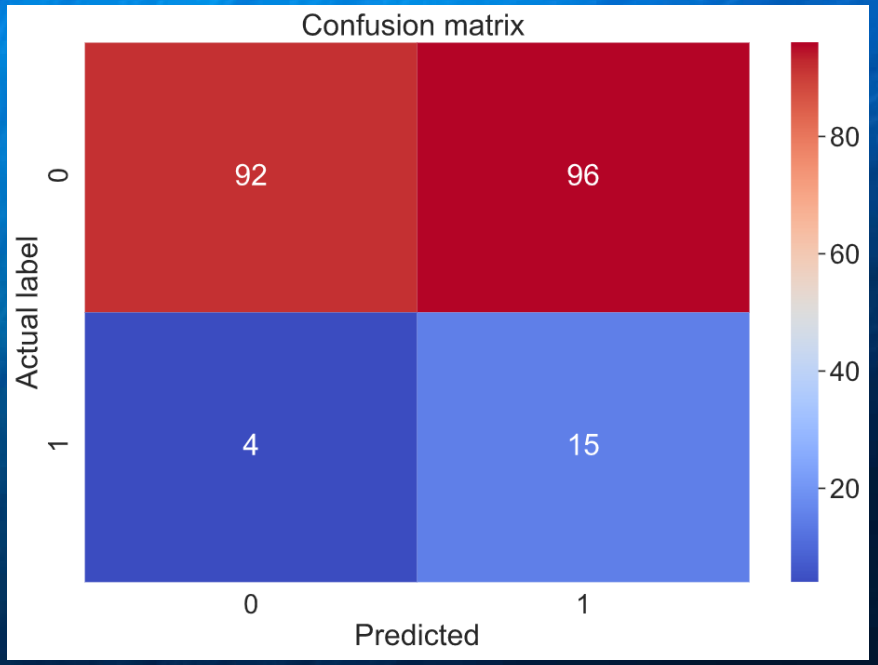


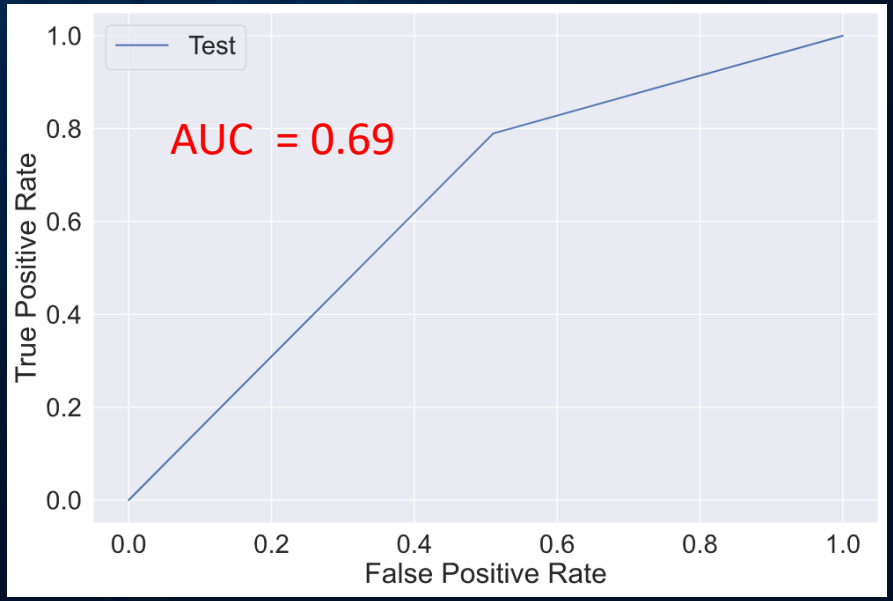
Loan Analysis – Decision Tree

Findings:

- As per decision tree model, dispo type whether an owner or user has taken a loan is the most important feature.
- Second important feature is credit card type.
- Finally, distric name plays an important role in loan defaults.



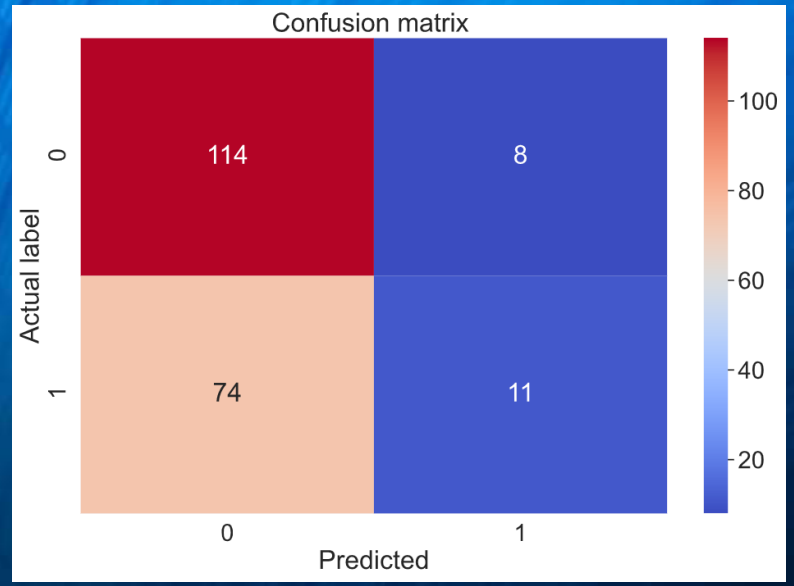


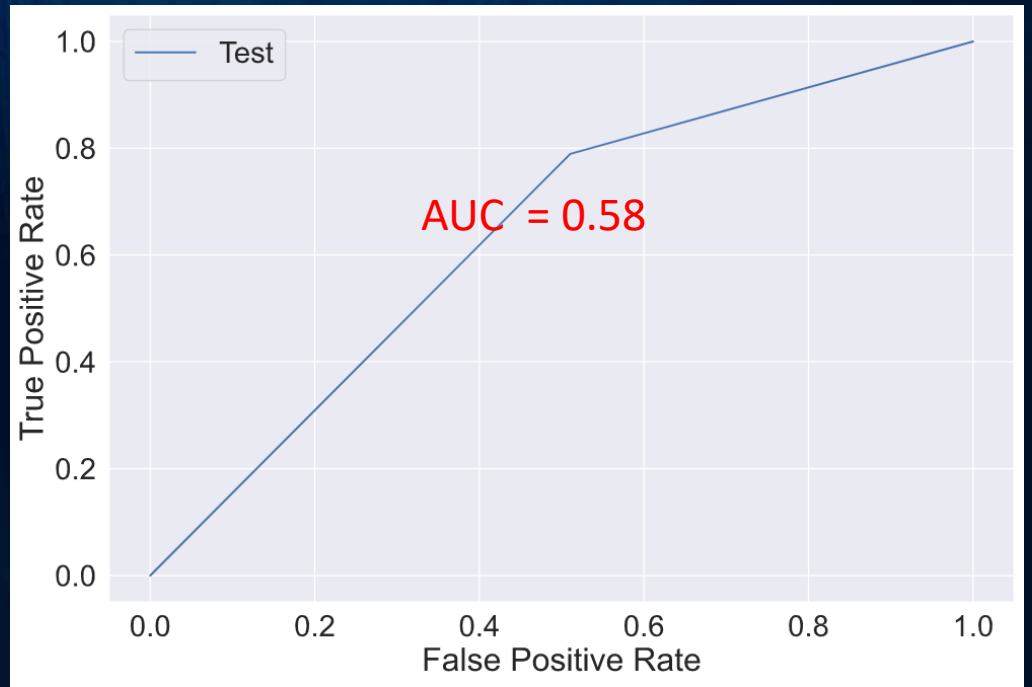


Loan Analysis – Logistic Regression with SMOTE

Findings:

- AUC of 0.58 makes this model very unreliable.
- This result is with SMOTE (balancing the dataset with artificial data points). Without SMOTE, AUC is much lower.





RESULTS

Here are the results of all the models ran on this dataset…

Models	Acc (Train)	Acc (Test)	F1 Score (Test)	AUC (Test)	Comment
Logistic Regression	0.90	0.90	0.95	0.45	
Random Forest	0.90	0.91	0.95	0.72	Best Model
Logistic Regression With SMOTE	0.99	0.99	0.74	0.58	
Random Forest With SMOTE	0.90	0.91	0.93	0.67	
Decision Tree	0.90	0.91	.095	0.63	

Summary of Analysis :-

1. There are 9% loans which are defaulted. This is extremely high compared with 1.02% (Jan 2020, S&P Consumer default Index)
2. There are 5300 clients while only 4500 account number indicating that many accounts have more than one client for example business account or family account.
3. There are certain district in which no. of defaulted loans are minimum or zero (further analysis is required)
4. There is a strong correlation between payment and duration of the loan (0.68) which is expected.
5. Loan originated in second half of 1995 have not defaulted due to some reason.
6. Total credit cards balances consists of 73% of classical card, 17% of gold members and 15% are juniors
7. Loan amount and total loan balance are the main features in loan default analysis.
8. There is no difference between no. of male and female defaulters, but seniors tend to have higher default rate.
9. Clients with loan between 100,000 and 200,000 have high default rate (slide #6)

Further Analysis

- Using feature engineering, more new features such a transactions per month/quarter, average salary per district etc to be included in analysis.
- Since data is from Czech Republic, domain expert in local financial market will be useful.
- New models such as XGBoost, Neural network etc. can be used to investigate performance improvement.

References

- **Dataset taken from - https://data.world/lpetrocelli/czech-financial-dataset-real-anonymized-transactions**
- **UNCC Coursework (ITCS 6265) - https://webpages.uncc.edu/mirsad/itcs6265/group1/index.html**
- **http://research.ganse.org/datasci/loanpredict/**


outperform
great
just
today
markets
price
rating
regal
new
will
like
capital
time
target
get
shares
amp

Thank You