Credit Card Approval Prediction

DSC630 Predictive Analytics T301(2235-1)

Kausik Chattapadhyay

Bellevue University

04/19/2023

Credit Card Approval Prediction

Kausik Chattapadhyay

Milestone 3 Preliminary Analysis

Data versus Expectations

The data I have chosen for this project should be able to answer all the questions I plan to respond to. The data I received in my first reference links was a lot larger than I had previously expected. It had a couple hundred thousand rows, which gave me a lot to play with. The data set had several categorical columns that helped denote whether the individual was getting approved or not, one thing I had to do was clean this up a bit. The number of features I wanted in my analysis I wanted to make sense, while I could choose all of them, it would make it difficult to get a successful model with them all. The data was plentiful and while I have a backup data set, I do not think I will need to use it as everything I need is in one.

Motivation

A Bank wants to automate the credit card eligibility process based on customer details provided while filing online application form & credit history of customer. They have given a problem to identify the customers segments which are eligible for Credit Card Approval, so that they can specifically target these customers. It is very important to manage credit risk and handle challenges efficiently for credit decision as it can have adverse effects on credit management. Therefore, evaluation of credit approval is significant before jumping to any granting decision. This project aims to produce such a model.

Furthermore, if the model proves robust enough, we could analyze which variables are the most predictive, or see if some combination of variables is predictive. That information could spark questions to inspire further research.

Visualizations

I am planning to use the below visualizations to explain my data:

1. **Correlation of data using Heatmap:** Heatmap will provide a correlation between every features of the dataset. This will provide data on which two features have the best correlation between them. There is no column (Feature) which is highly co-related with 'Status'.
2. **Pie Graph/Histogram/Scatterplot:** Percentages of Applications approved based on Gender, owning a car, owning a real state property, children count, income. These will give some insights like majority of application are approved for Female's, majority of applicant's dont own a car, majority of applicant's own a Real Estate property / House, majority of applicant's don't have any children.
3. **Pie Graph:** Percentages of Applications submitted based on Income Type shows that, majority of applicant's are working professional. Percentages of Applications submitted based on education shows majority of applicant's completed the Secondary Education.
4. **Bar Plot:** Number of Applications submitted based on family status shows majority of applicant's are married. Number of Applications submitted based on Housing Type shows majority of applicant's lives in House / Apartment.
5. **Histogram:** This graph shows that, majority of applicant's are 25 to 65 years old, majority of applicant's are Employed for 0 to 7 years.

6. **Scatter Plot:** These graph shows majority of applications are rejected if Total income & years of Employment is less, majority of applications might be rejected if Total income is less, Age based rejection is equally distributed

Some of the visualizations I will use are some bar charts to show the number of counts in comparison of each feature as well as the approval and denial rate. These charts are great for initial EDA exploration to understand the data and find what variables I will be using. I am planning on some scatter charts as well to show relationships between features.

Model versus expectations.

Based on my experience, there will be an imbalance in the data. To rebalance the data, I am planning to use SMOTE (Synthetic Minority Oversampling Technique) technique. As there are multiple datasets and observations, we will most likely experience imbalanced data.

Apart from this, we will have data cleansing, which needs to be done as the data we get will be less required.

1. Eliminating duplicates
2. Replace NA or Null values with Median values.
3. Categorical data which has extreme values need to be replaced with proper values

As per the initial data analysis, Logistic Regression and Decision Tree should work. However, I am also planning to implement Random Forrest if time permits. Implementing more than one model will benefit us by comparing them based on accuracy score and choosing the best model for this dataset.

I am confident that Logistic Regression and Decision Tree will work, and there is no need to replace them with other models.

I am confident that my original expectations are still reasonable as I have done the initial groundwork to come up with these choices. However, I have also made backup plans to help in the worst-case scenario. I might be using only some of the datasets as they might have some redundant data. After merging the datasets, we might have to remove the duplicates.

On a positive note, selected data and models will help us build decent models suitable for the prediction.

I had written that I was going to use a linear regression in my week 1 milestone, but this was incorrect as I will be looking at features and the target with be a Boolean approval value. This led me down the path of using a logistic regression to help me solve the data. The data set being quite large allows me to parse out a small training data set. I ran the model on 1 training data set and I did get data showing that it showed some sense of accuracy. I plan to make another training data set to see if my accuracy is still the same, or maybe try to train on a larger data group, my initial training split was only 80/20.

I think I might want to explore a few different features from the original few, like income, marital status, home or car ownership, that I had played with. With some different features that I bring into the mix I may get varying results and I will need to play around with several options.

Motivation and expectations

I think the data is solid and will allow me to get everything I need. My original expectations of being able to calculate loan approval rates seem on track. I did mention I wanted to see if there were differences in different areas/populations when I started this project. My data set does not seem to have the location or ethnicity factored into it, as those things shouldn't matter, but it does

have gender. That doesn't seem like the strongest lever to pull so I may try to do some research and pull from other scientific models to get examples of other people and how they approached that question. This can lead to a future study or help fill in some gaps when presenting the data.

Credit Card Approval Prediction

Kausik Chattapadhyay

Milestone 2 Project Proposal

I am planning on Credit Card Approval Prediction. A Bank wants to automate the credit card eligibility process based on customer details provided while filing online application form & credit history of customer. They have given a problem to identify the customers segments which are eligible for Credit Card Approval, so that they can specifically target these customers.

The decision of approving a credit card or loan is majorly dependent on the personal and financial background of the applicant. Factors like age, gender, income, employment status, credit history and other attributes all carry weight in the approval decision. Credit analysis involves the measure to investigate the probability of a third party to pay back the loan to the bank on time and predict its default characteristic. Analysis focus on recognizing, assessing, and reducing the financial or other risks that could lead to loss involved in the transaction.

There are two basic risks: one is a business loss that results from not approving the good candidate, and the other is the financial loss that results from by approving the candidate who is at bad risk. It is very important to manage credit risk and handle challenges efficiently for credit decision as it can have adverse effects on credit management. Therefore, evaluation of credit approval is significant before jumping to any granting decision.

Dataset

For this project, I will be using datasets from Kaggle.com. Data Sets: Two csv files – Application Record and Credit Record. Data Variables are age, gender, income, education, years employed, credit history, months balance, status (default payment of not) among others.

https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction

https://www.kaggle.com/code/rikdifos/credit-card-approval-prediction-using-ml/input

https://www.kaggle.com/datasets/caesarmario/application-data

https://www.kaggle.com/code/itsual/credit-card-approval-e2e/input

Variable descriptions are below:

Feature name	Explanation
ID	Client number
CODE_GENDER	Gender
FLAG_OWN_CAR	Is there a car
FLAG_OWN_REALTY	Is there a property
CNT_CHILDREN	Number of children
AMT_INCOME_TOTAL	Annual income
NAME_INCOME_TYPE	Income category
NAME_EDUCATION_TYPE	Education level
NAME_FAMILY_STATUS	Marital status
NAME_HOUSING_TYPE	Way of living
DAYS_BIRTH	Birthday
DAYS_EMPLOYED	Start date of employment
FLAG_MOBIL	Is there a mobile phone
FLAG_WORK_PHONE	Is there a work phone
FLAG_PHONE	Is there a phone
FLAG_EMAIL	Is there an email
OCCUPATION_TYPE	Occupation
CNT_FAM_MEMBERS	Family size

Feature name	Explanation
ID	Client number
MONTHS_BALANCE	Record month
STATUS	Status

Status 0 indicates that the applicant has paid their credit due on time or has no loans remaining. Whereas 1 indicates that they are behind on payments.

What type of Models do you plan to use and Why?

I am planning to use the below models:

1. **Logistic Regression:** This model is used to predict if a given credit card or loan application got approved or not using the various attributes provided in the dataset. Logistic regression analysis is valuable for predicting the likelihood of an event. It helps determine the probabilities between two classes. In a nutshell, by looking at historical data, logistic regression can predict whether: A credit card application is approved or not. I prefer Logistic over Linear because of the binary classification problem while linear regression is used for predicting numeric values.
2. **Decision Tree Implementation:** This model is a tree-based classification model. With the available attributes, it will predict the target variable, which is credit card application approved or not, using various independent variables. Decision Trees bisect the space into smaller and smaller regions, whereas Logistic Regression fits a single line to divide the space exactly into two. Of course, for higher-dimensional data, these lines would generalize to planes and hyperplanes. I prefer Decision Tree since it's simple to understand and to interpret and require little data preparation, able to handle both numerical and categorical data and able to handle multi-output problems.
3. **Random Forest:** Whether I have a regression or classification task, random forest is an applicable model for my needs. It can handle binary features(special cases of categorical), categorical features, and numerical features. There is very little pre-

processing that needs to be done. The data does not need to be rescaled or transformed. They are parallelizable, meaning that I can split the process to multiple machines to run. This results in faster computation time. Boosted models are sequential in contrast, and would take longer to compute. Random forests are great with high dimensional data since I am working with subsets of data. It is faster to train than decision trees because I am working only on a subset of features in this model, so I can easily work with hundreds of features. Prediction speed is significantly faster than training speed because I can save generated forests for future uses. Random forest handles outliers by essentially binning them. It is also indifferent to numeric or categorical features, handles unbalanced data. Each decision tree has a high variance, but low bias. But because I average all the trees in random forest, I am balancing out the potential biases and reducing the variance.

4. I would try all other models like **Support Vector Machines, K Nearest Neighbors, XgBoost.** Most of the Kaggle competitions won with Xgboost. XGBoost is designed to handle missing values internally. The missing values are treated in such a manner that if there exists any trend in missing values, it is captured by the model. User is required to supply a different value than other observations and pass that as a parameter. XGBoost tries different things as it encounters a missing value on each node and learns which path to take for missing values in future. SVM does not perform well with missing data. It is always better to impute the missing values before running SVM.

   Naive Bayes does not perform well with data scarcity. For any possible value of a feature, you need to estimate a likelihood value by a frequentist approach. This can

result in probabilities going towards 0 or 1, which in turn leads to numerical instabilities and worse results.

How do you plan to evaluate your results?

I will use an accuracy score with K-Fold cross validation, AUC score and confusion matrix to evaluate the results using a train test data split (80-20) which I have been doing so far. I will choose the model based on this accuracy.

What do you hope to learn?

Using this analysis, I am hoping to learn what are the major factors which would cause Credit card or loan approval. Also, what are the reasons of most credit card prediction as denial? If the candidates have lower income and less years of employment, is he/she a good candidate? What's the good candidate's demographic data? (1. Is he/she married? 2. Do they live in apartments/houses? 3. What's the age limit? 4. How many years are they employed?) When I know these details, it can help Banks mitigate risks. Credit analysis involves the measure to investigate the probability of a third-party to pay back the loan to the bank on time and predict its default characteristic. Analysis focus on recognizing, assessing, and reducing the financial or other risks that could lead to loss involved in the transaction.

There are two basic risks: one is a business loss that results from not approving the good candidate, and the other is the financial loss that results from by approving the candidate who is at bad risk. It is very important to manage credit risk and handle challenges efficiently for credit decision as it can have adverse effects on credit management. Therefore, evaluation of credit approval is significant before jumping to any granting decision.

Assess any risks and ethical implications with your proposal:

1. As the data is sourced from a public website and not a government departmental source or actual bank data, I am unsure if the information is accurate.

2. This is not recent data, and as research evolves and might have added a few more attributes, I might be missing those extra parameters. However, I will be researching more datasets to get the latest data.

Identify a contingency plan if your original project plan does not work out:

This dataset should suffice with the initial research done for finalizing the dataset. However, I will look for the most recent data, primarily from the sources like Bank Masked data, to see if that will add more value to the analysis.

In case of any data imbalance issues, I will use techniques like SMOTE to make the prediction accurate.

If the Credit card or Loan data is not helping to proceed further, I have another contingency plan to use Online fraud detection datasets.

Include anything else you believe is important:

Apart from all the details stated above, I will also look for any other ways I have to evaluate the results. I will perform splitting datasets Train (80%) and Test (20%) and took 2 passes at the Machine learning models, one with initial data and other with balanced data after performing SMOTE technique. I will impute if any missing values or drop the field if not required. I will

perform several EDA's to find out some questions like What's the good candidate's demographic data? (1. Is he/she married? 2. Do they live in apartments/houses? 3. What's the age limit? 4. How many years are they employed?). Apart from the above-mentioned models, I will try other models from deep learning also to understand if I can get better accuracy.

Also I have to figure out how to improve accuracy score is confusion matrix and accuracy score is showing low and check for underfitting and overfitting scenarios.

References

https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction

https://www.kaggle.com/code/rikdifos/credit-card-approval-prediction-using-ml/input

https://www.kaggle.com/datasets/caesarmario/application-data

https://www.kaggle.com/code/itsual/credit-card-approval-e2e/input

https://medium.com/@taniyaghosh29/machine-learning-algorithms-what-are-the-differences-9b71df4f248f

https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222

https://towardsdatascience.com/15-must-know-machine-learning-algorithms-44faf6bc758e

https://www.quora.com/Why-does-XGBoost-perform-better-than-SVM

https://arxiv.org/pdf/2009.06366.pdf

https://dibyendudeb.com/comparing-machine-learning-algorithms/