Week 5 Case Study- Prediction of Mortgage Risks

Professor: Dr. Fadi Alsaleem

Predictive Analytics

Author: KARTHIKEYAN CHELLAMUTHU

Due Date: 08/10/2022

Table of Contents

Introduction

What was the problem being solved?.....3

Why was this problem important to solve?.....3

How was the data acquired?.....3

Methods and Results

What steps were taken to prepare the data? 6

How was this problem solved? 7

What modeling techniques were used? 8

What metrics were used to evaluate the results? Why was this metric chosen?.....10

Conclusion

How were the results or model implemented? 10

What were the actionable consequences of the case study?.....11

What did the team learn from the case study? 12

How should or would the team approach the problem differently in the future?.....12

Reference 13

Introduction

- **What was the problem being solved?**

  Lending Club company which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. There are risks are associated with the bank's decision.

- **Why was this problem important to solve?**

  It is important for an evaluation study for example, if he or she is likely to default, then approving the loan may lead to a financial loss for the company. Solving this case study will give us an idea about how real business problems are solved using EDA and Machine Learning. In this case study, we will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

- **How was the data acquired?**

  The data can be downloaded from the Kaggle using the below link, Here is some more information about the data set

  https://www.kaggle.com/code/faressayah/lending-club-loan-defaulters-prediction?scriptVersionId=74775795&cellId=4

Loan Stat New	Description
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
int_rate	Interest Rate on the loan
installment	The monthly payment owed by the borrower if the loan originates.
grade	LC assigned loan grade
sub_grade	LC assigned loan subgrade
emp_title	The job title supplied by the Borrower when applying for the loan.*
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
annual_inc	The self-reported annual income provided by the borrower during registration.
verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified

Loan Stat New	Description
issue_d	The month which the loan was funded
loan_status	Current status of the loan
purpose	A category provided by the borrower for the loan request.
title	The loan title provided by the borrower
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.
addr_state	The state provided by the borrower in the loan application
dti	A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income.
earliest_cr_line	The month the borrower's earliest reported credit line was opened
open_acc	The number of open credit lines in the borrower's credit file.
pub_rec	Number of derogatory public records
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.

Loan Stat New	Description
total_acc	The total number of credit lines currently in the borrower's credit file
initial_list_status	The initial listing status of the loan. Possible values are – W, F
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
mort_acc	Number of mortgage accounts.
pub_rec_bankruptcies	Number of public record bankruptcies

Methods and Results

- **What steps were taken to prepare the data?**

  The data collected for this project having numerical and categorical data types. The target variable here 'loan status' is also categorical data. To prepare data for modeling stage, we need to transform all the features into numerical.

  Missing values, since data collected from historical and there will be null for customers who do not have accounts with the financial institutions or Lending Club company. To fix this issue, the data is filled with values from the application when customer applied for the loan.

  In addition, author has performed the following Exploratory Data Analysis and Cleaning exercise to prepare the data for modelling.

  - Remove or fill any missing data.

- Remove unnecessary or repetitive features.
- Convert categorical string features to dummy variables.
- Correlations and Multicollinearity
- EDA & Process Train Set
- Categoricals
- Continuous

- **How was this problem solved?**

  This case study has brought multiple attributes to be evaluated. It was not easy to ignore these factors to solve the problem. Then, author used normalization technique to choose the risk types which make up most of the dataset. Since every individual has different type of loan history in the part all those factors are the driving force to identify the same. Duplicate and missing values checks have also been performed and found neither duplicates nor missing values.

- Current status of the loan
- Installment: The monthly payment owed by the borrower if the loan originates.
- loan_amnt: The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
- Grade: LC assigned loan grade
- Sub_grade: LC assigned loan subgrade
- term: The number of payments on the loan. Values are in months and can be either 36 or 60.

- Home_ownership: The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
- Verification_status: Indicates if income was verified by LC, not verified, or if the income source was verified
- Purpose: A category provided by the borrower for the loan request.
- Int_rate: Interest Rate on the loan
- Annual_inc: The self-reported annual income provided by the borrower during registration
- Emp_title: The job title supplied by the Borrower when applying for the loan.
- Emp_length: Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
- It seems that loans with high intersest rate are more likely to be unpaid.
- Only 75 (less then) borrowers have an annual income more than 1 million, and 4077
- Title: The loan title provided by the borrower
- Pub_rec: Number of derogatory public records
- Initial_list_status: The initial listing status of the loan. Possible values are – W, F
- Application_type: Indicates whether the loan is an individual application or a joint application with two co-borrowers
- Mort_acc: Number of mortgage accounts
- Pub_rec_bankruptcies: Number of public record bankruptcies

- **What modeling techniques were used?**

    - Artificial Neural Networks (ANNs)
    - XGBoost Classifier

- Random Forest Classifier

The first model used is ANN, as we know Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (SNNs), are a subset of machine learning and are at the heart of deep learning algorithms. Their name and structure are inspired by the human brain, mimicking the way that biological neurons signal to one another.

The second algorithm used to train model is XGBoost, the XGBoost algorithm is effective for a wide range of regression and classification predictive modeling problems. It is an efficient implementation of the stochastic gradient boosting algorithm and offers a range of hyperparameters that give fine-grained control over the model training procedure. Since predicting customer will be default or close the loan is a type of classification problem so this algorithm should be best fit for this model.

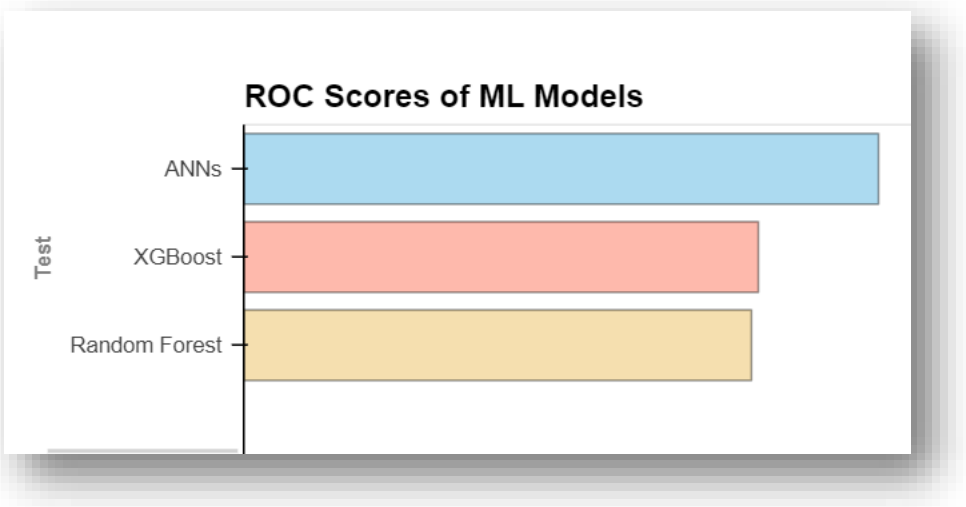
The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

After training data on all three algorithms, the models are tested using test data and predictions are made, then accuracy score is calculated to compare the models. The accuracy score tells us which model can predict a most number of target variable correctly.

Based on the comparison ANN model accuracy score is better than other two models.

Model trained using ANN algorithm is the baseline model and will be used in prediction model.

Compared the model performance at last to get best out of it.



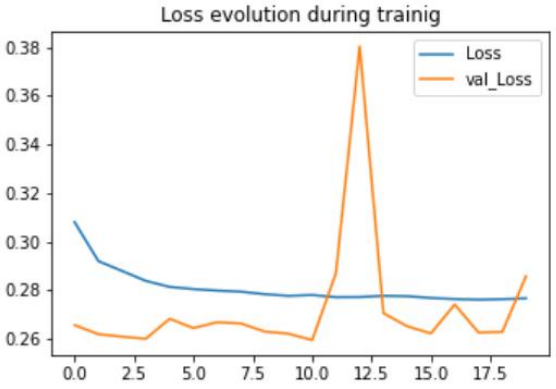
- **What metrics were used to evaluate the results? Why was this metric chosen?**

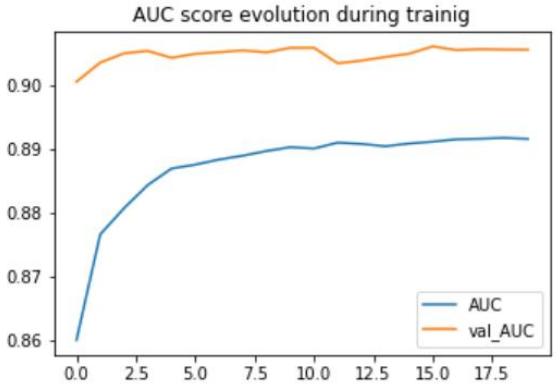
- A confusion matrix -is a table that is used to define the performance of a classification algorithm. A confusion matrix visualizes and summarizes the performance of a classification algorithm.
- MAE - The Mean absolute error represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset.
- "Extreme Gradient Boosting" and it is an implementation of gradient boosting trees algorithm.
- RMSE - Root Mean Squared Error is the square root of Mean Squared error. It measures the standard deviation of residuals.

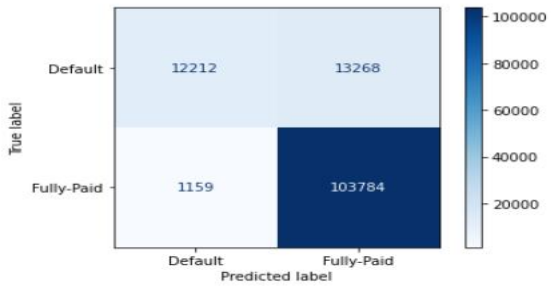
Conclusion

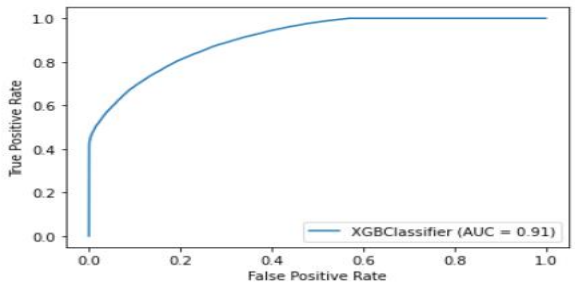
- **How were the results or model implemented?**

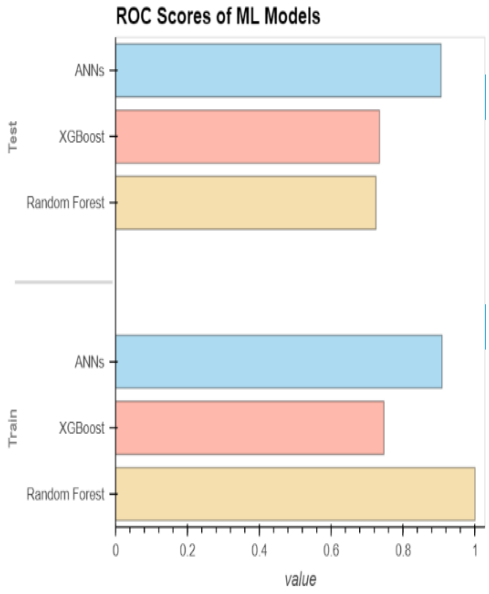
Author has tried to improve the result by running various model such as Ann and following is the output and plots for those models were the results were published.











- **What were the actionable consequences of the case study?**

Two types of risks are associated with the bank's decision:

If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company. If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

The data given contains the information about past loan applicants and whether they 'defaulted' or not. The aim is to identify patterns which indicate if a person is likely to default, which may be used for takin actions such as denying the loan, reducing the amount of loan, lending (too risky applicants) at a higher interest rate, etc.

- **What did the team learn from the case study?**

  Solving this case study will give us an idea about how real business problems are solved using EDA and Machine Learning. In this case study, we will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimize the risk of losing money while lending to customers.

- **How should or would the team approach the problem differently in the future?**

  When a person applies for a loan, there are two types of decisions that could be taken by the company to take the pre-requisite based on the above study.

  1. Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:
     - Fully paid: Applicant has fully paid the loan (the principal and the interest rate)
     - Current: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.

- Charged-off: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan

2. Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

References:

https://www.kaggle.com/code/faressayah/lending-club-loan-defaulters-prediction?scriptVersionId=74775795&cellId=4