Loan Default Prediction

Insight Miners - Monash University - Team 3

Problem Statement

Understanding whether to lend a borrower money or not, helps with the mitigation of credit risk. Our objective is to:

- Evaluate potential clients to determine whether they are likely to default or not
- Determine factors which influence credit risk.

Hypothesis

“The probability of default, sometimes abbreviated as POD or PD, expresses the likelihood the borrower will not maintain the financial capability to make scheduled debt payments. For individual borrowers, default probability is most represented as a combination of two factors: debt-to-income ratio and credit score.Credit rating agencies estimate the probability of default for businesses and entities that issue debt instruments, such as corporate bonds. Generally speaking, higher PODs correspond with higher interest rates and higher required down payments on a loan.” - *What Factors Determine Credit Risk? (2021, August 12). Investopedia. https://www.investopedia.com/ask/answers/022415/what-factors-are-taken-account-quantify-credit-risk.asp*

The aim is to detect the expected and discover the unexpected

Existing Data

- 47 columns/variables describing the characteristics of the borrower
- 'isDefault' is a boolean column that determines whether a client has defaulted or not, which we are trying to predict.

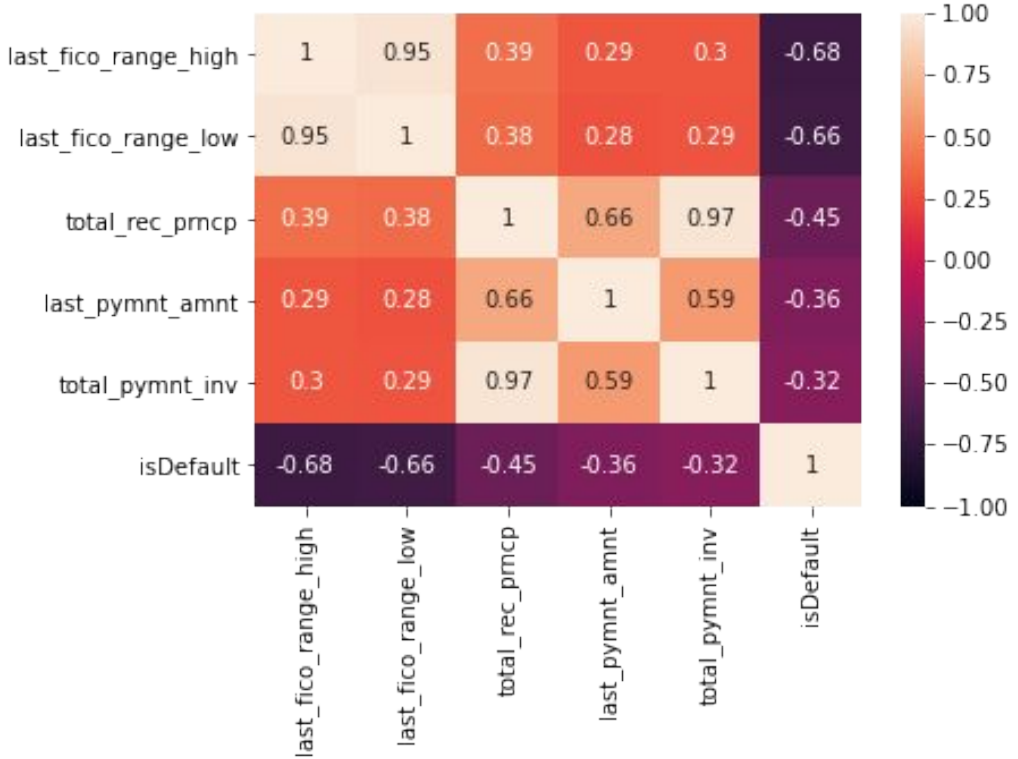
	id	loanAmnt	term	interestRate	installment	grade	subGrade	employmentTitle	employmentLength	homeOwnership	isDefault
0	0	35000.0	5	19.52	917.97	E	E2	320.0	2 years	2	1
1	1	18000.0	5	18.49	461.90	D	D2	219843.0	5 years	0	0
2	2	12000.0	5	16.99	298.17	D	D3	31698.0	8 years	0	0
3	3	11000.0	3	7.26	340.96	A	A4	46854.0	10+ years	1	0
4	4	3000.0	3	12.99	101.07	C	C2	54.0	NaN	1	0

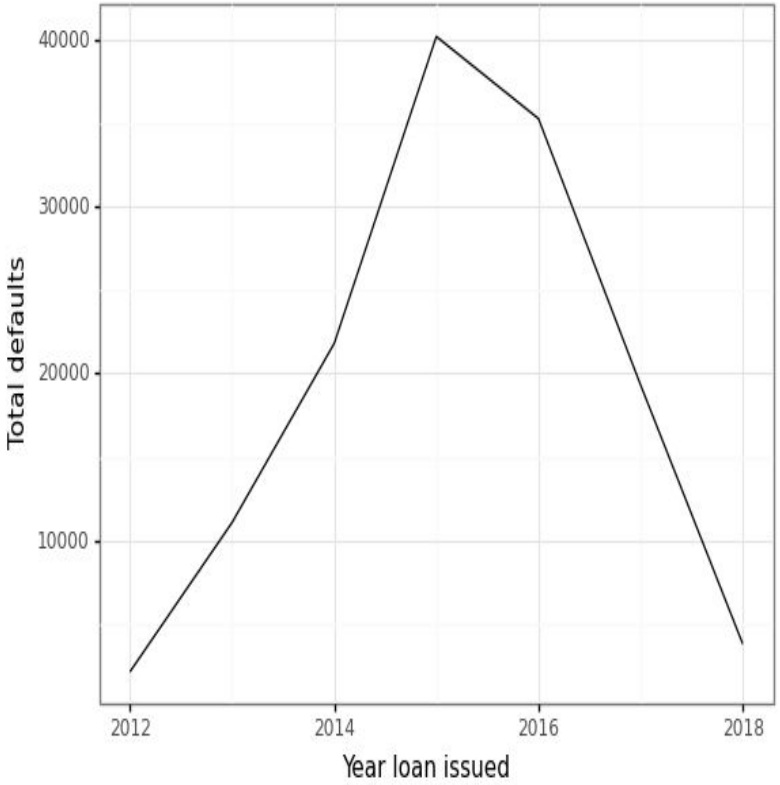
Analysis

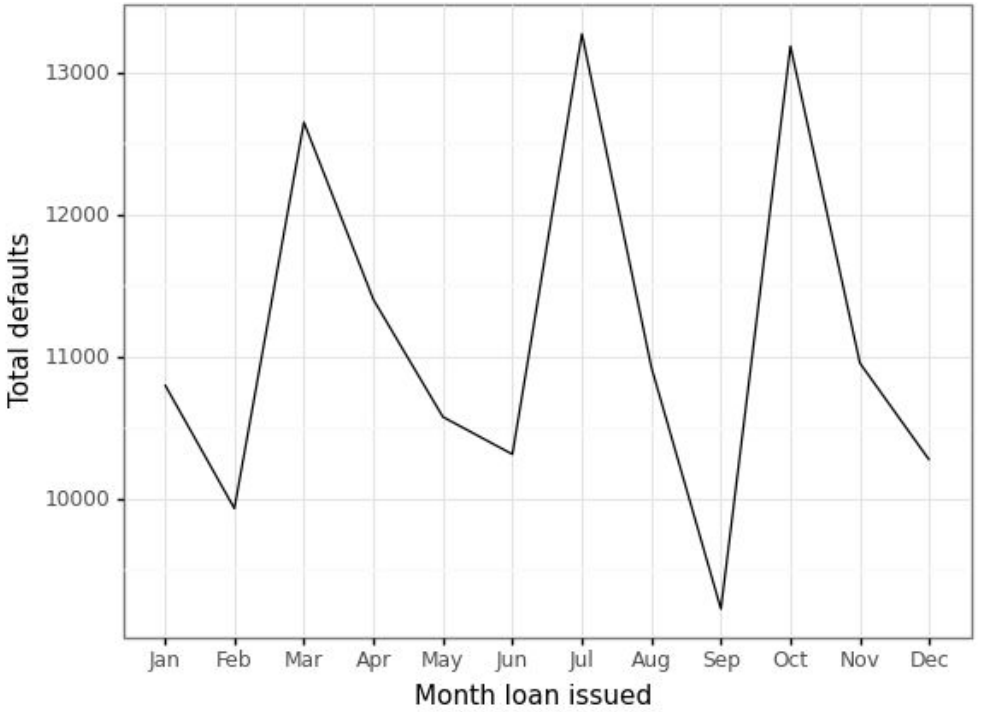
What insights did we generate?

Correlation Heatmap

A pearson correlation denotes the strength of linear correlation between two variables. The higher the magnitude(positive or negative), the more associated the two variables are. Take particular notice of the purple cells.



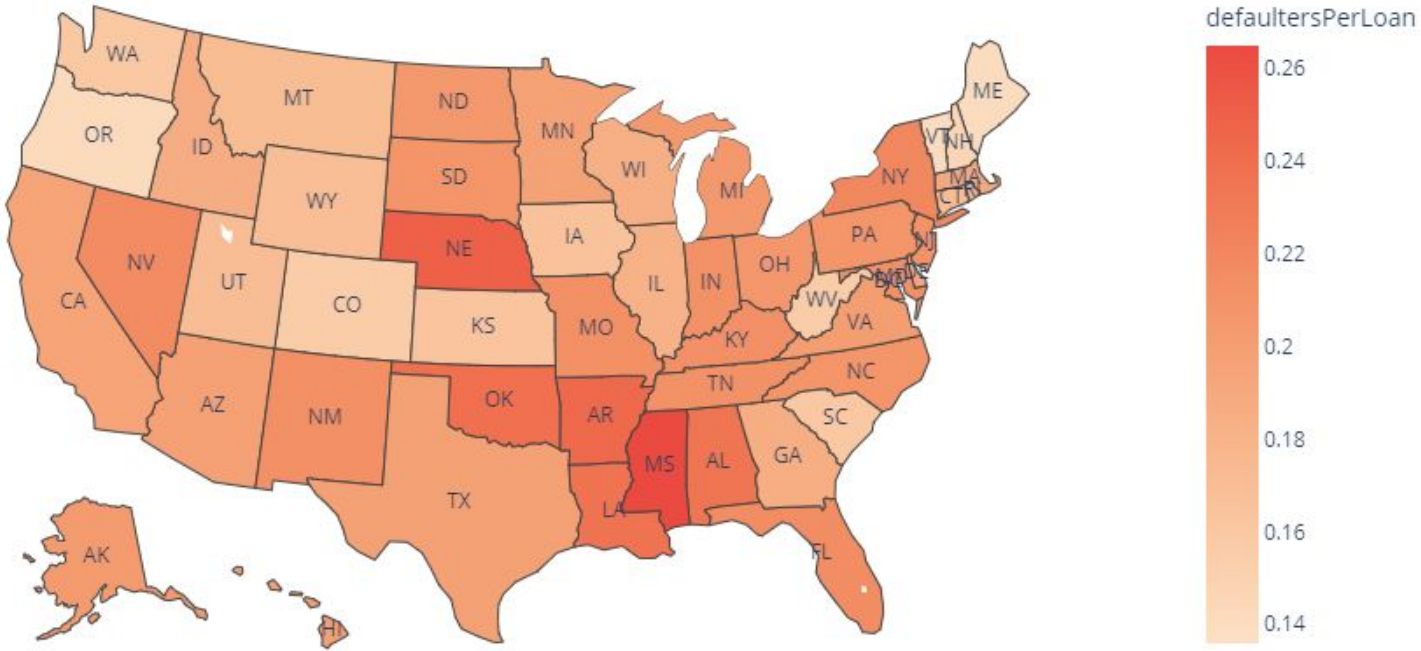




Analyzing the trend in number of defaulted loans across years and months

Time-series analysis of defaulters

- From the yearly distribution of total defaulters graph, it is quite evident that the number of defaulters was the lowest in 2012 and and highest in 2015.
- From the monthly distribution of total number of defaulters, we can detect seasonality with three prominent peaks occurring in the months of March, July and October.
- September has the lowest number of defaulters by a large margin.



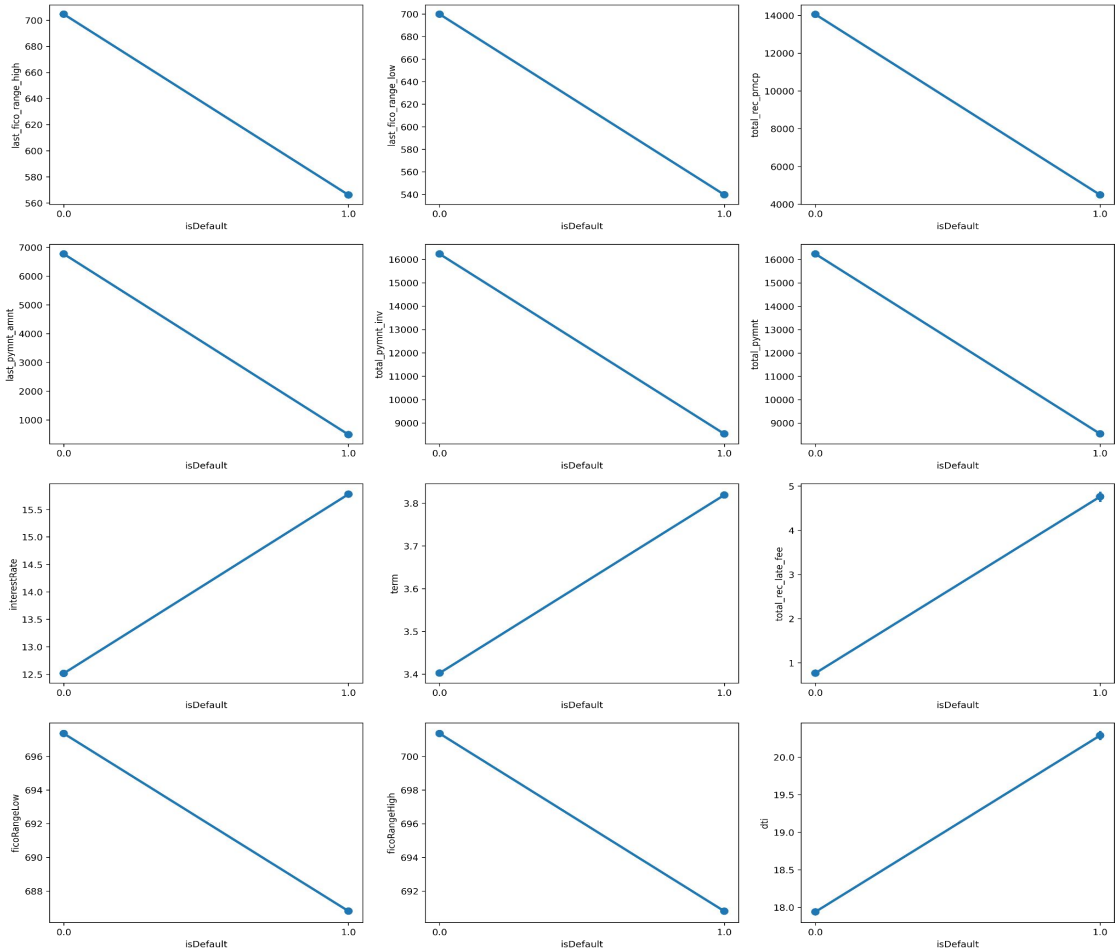
Using the spatial distribution of defaulters per total count of loans per state to observe the states with highest proportion of defaulters

State-wise distribution of total defaulters

- The map depicts the ratio of defaulters to the total number of loans issued for each state.
- With reference to the map on the previous slide, we can conclude that certain states such as Mississippi and Nebraska have the most defaulters to loaners ratio.
- On the contrary, states like Maine, Vermont and New Hampshire had the lowest defaulters to loaners ratio.

Behaviour of Defaulters vs Non-Defaulters

Vis 3



Behaviour of Defaulters vs Non-Defaulters

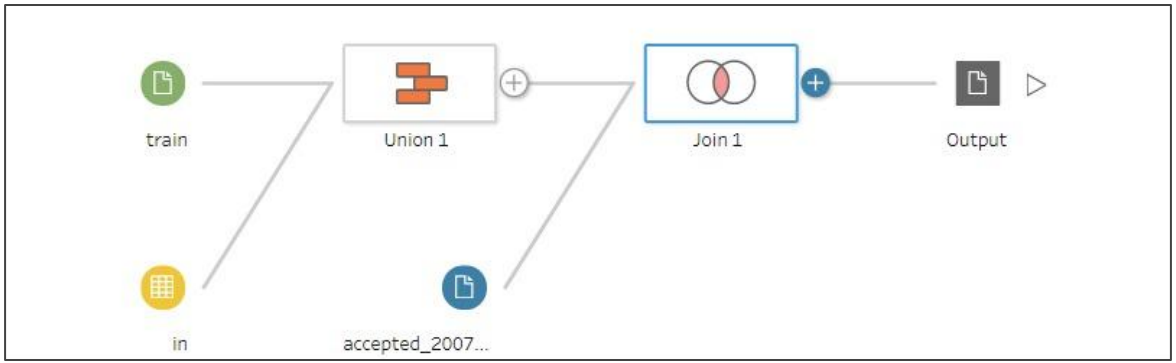
- Across all the plots on the previous slide, we can observe a stark difference between mean values of variables for defaulting and non-defaulting borrower
- Notice how for non-defaulters the mean values of the following variables are high: last_fico_range_high, last_fico_range_low, total_rec_prncp, last_pymnt_amnt, total_pymnt_inv, total_pymnt, fico_range_high and fico_range_low
- On the contrary, notice how for defaulters the mean values of the following variables are high: interest rate, term, total_rec_late_fee(total late fees received) and dti(debt-to-income)

Methodology

Data to Insights

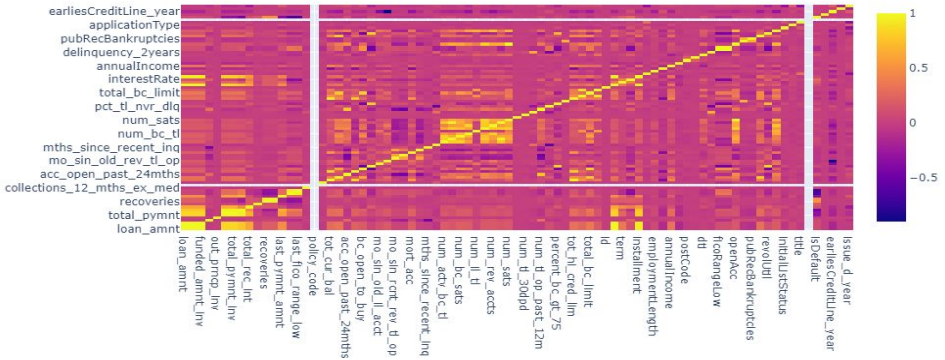
The journey of data to insights and prediction is summarized below

01. Reading the given datasets
02. Merging the additional dataset with given one



03. Performing data cleaning by trimming or imputing missing data and also handling outliers and inaccuracies in the dataset.

04. Performed correlation analysis using the Pearson correlation heatmap and Chai square test, to identify features correlated with “isDefault”



Correlation Heatmap - extremes imply association

05. Preparing the dataset for modelling by splitting it into features and target variable and performing K stratified sampling to account for imbalance in class ratios

06. Training different machine learning models on the training dataset and evaluating the model based on ROC-AUC score

Supplementary Research

- ➔ We explored additional datasets in search of additional factors that could potentially influence the probability of defaulting and also help us better understand the behaviour of defaulters.
- ➔ The search resulted in a dataset to be found with similar features to the given dataset in Kaggle -*All Lending Club loan data*. (2019, April 10). Kaggle. https://www.kaggle.com/wordsforthewise/lending-club
- ➔ The two datasets were joined based on multiple keys forming a primary key and then the additional features were included as part of the analysis and prediction

Key Results

Factors to be considered for identifying default risk based on experimentation

- Debt-to-income ratio
- Historical trends in default rates
- Loan term
- FICO (credit) Score

Based on the ANOVA F-test which tests if means of groups significantly differ from each other against isDefault target variable, we found the F scores described in the table on the right.

NOTE: Lower the F-Score, more similar are the sample/variable means, more the influence of the variable to POD(Probability of Default)

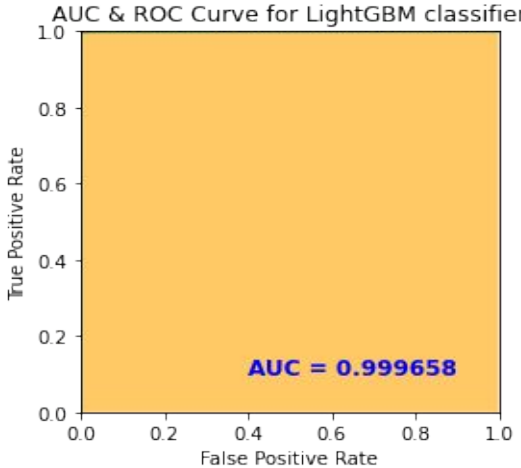
Upper Limit Range of FICO Score	590566.713
Lower Limit Range of FICO Score	525862.373
Principle received to date	170042.852
Last Payment Amount	99340.740
Number of payments received	75822.225
Total Payment	75821.840
Interest Rate	54145.597

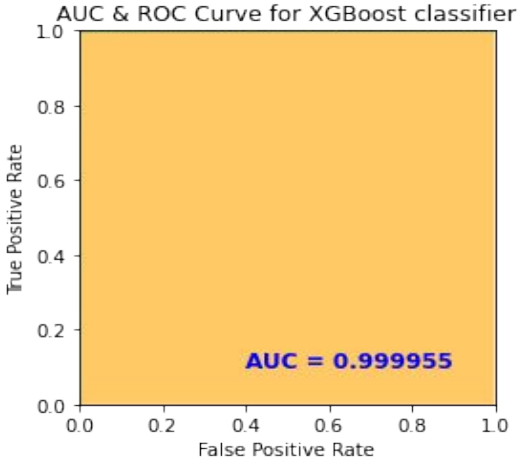
Inferences

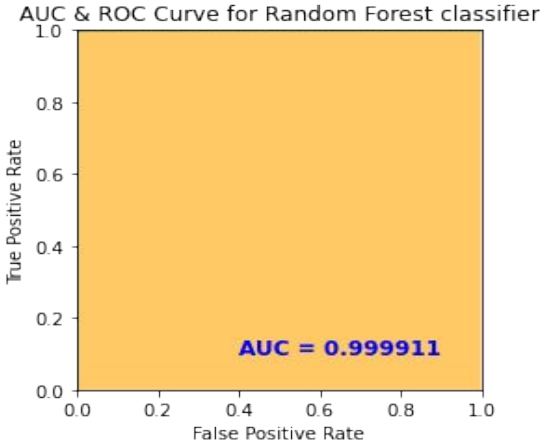
- In our initial hypothesis, we stated that probability an individual borrower defaults is directly proportional to the interest rate and required down payment on the loan. Adequate stress was put on the importance of the individual's debt-to-income ratio and their credit score.
- The last recorded FICO score of the individual proved to be the most integral factor of them all, receiving the highest F-score by a large margin.
- The FICO score recorded at issue date received a much lower F-score compared to the last recorded FICO. This is possibly because the last recorded FICO is a much better indicator of the borrower's most recent credit score, thus influencing the probability of default by a larger margin.
- The debt-to-income ratio had an underwhelming performance compared to the other factors mentioned in our hypothesis.

Comparing model performance

We compared the performance the performance metrics of XGBoost, LightGBM, CatBoost and RandomForest Classifiers and found that LightGBM was ahead than others in terms of training time but not accuracy, yet we decided to go with it as the scores for other models were only marginally better(decimals upto the 4th place). The LightGBM classifier was trained on features chosen based on ANOVA F test described earlier and we found the “Area under Curve - Receiver operating characteristic”(AUC-ROC) to be close to 0.99







Issues faced/Possible flaws

- The merging of additional dataset and the given dataset was complex due to the size of the both of the datasets
- Merging resulted in further effort in cleaning the dataset
- There is a possibility that fractionally, bias might have been introduced due to imputation of missing values
- The removal of outliers may have resulted in loss of useful information

Conclusion

Conclusion

With the assistance of statistical methods and machine learning, we have not only affirmed our initial hypothesis, but also found other key features which are highly influential with the probability of default. Not only that, we were able to profile and differentiate defaulters from the non-defaulters

THANK YOU