Segmenting and Predicting Loan Repayment Probability of Lending Club Debtors

Our Team

Uswa Mazhar

Sakshi Joshi

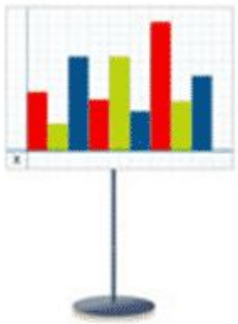
Jose Repettoparedes

Shakthi Viswanathan

What is Lending Club?

- LendingClub is an American peer-to-peer lending company, headquartered in San Francisco, California.
- It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market.
- At its height, LendingClub was the world's largest peer-to-peer lending platform

How Lending Club Works

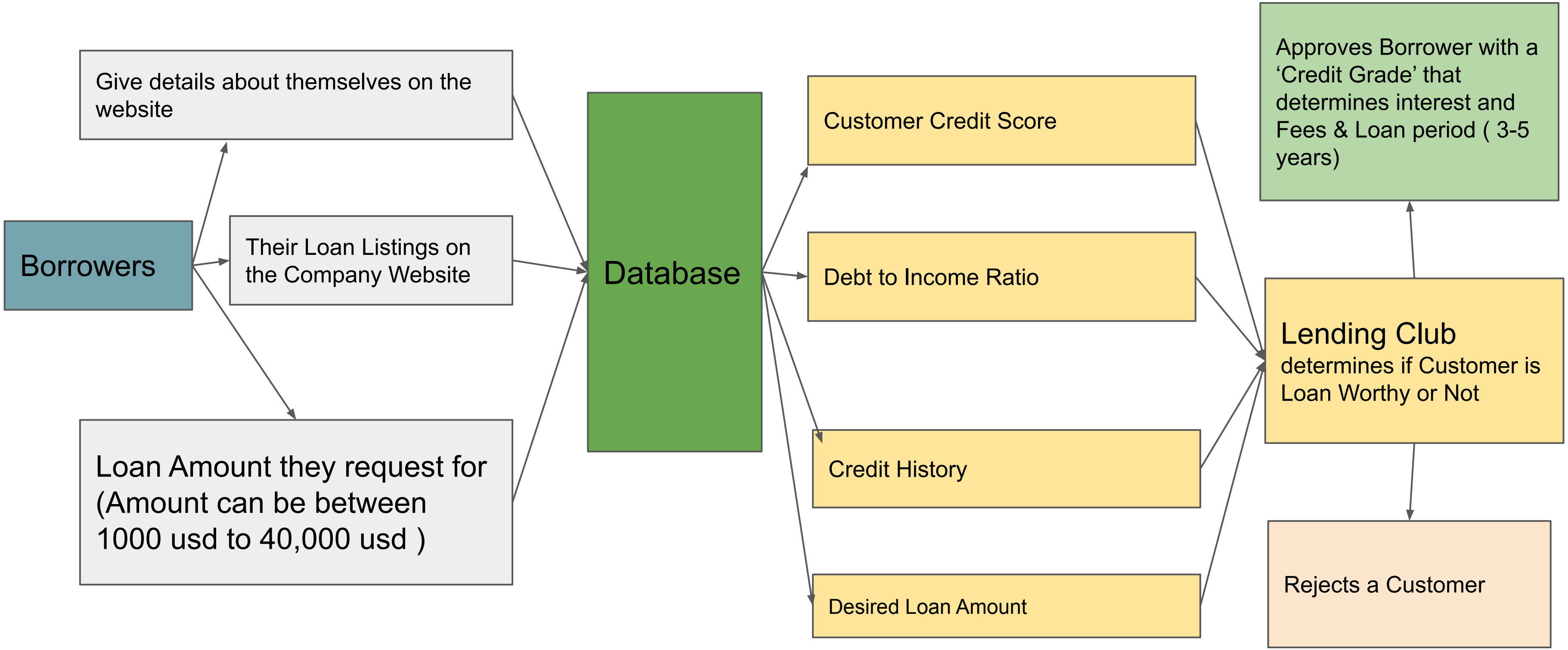


Borrowers apply for loans.
Investors open an account.

Borrowers get funded.
Investors build a portfolio.

Borrowers repay automatically.
Investors earn & reinvest.

Detailed Working of The Lending Club



Project Goal

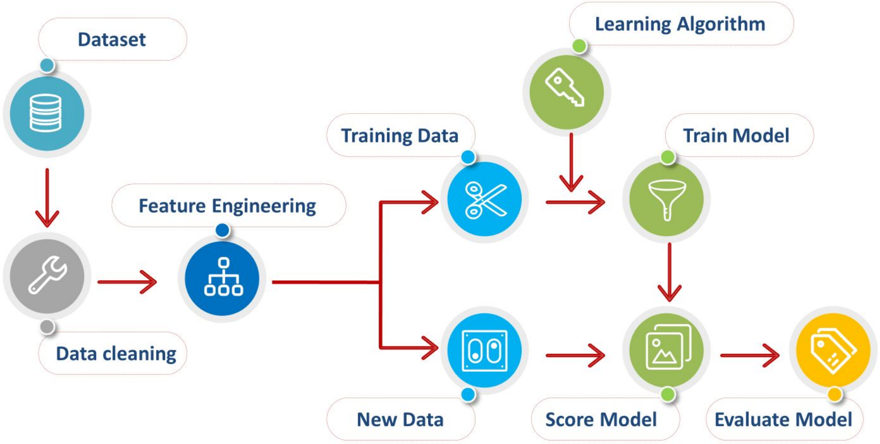
LendingClub

The goal of the project is the analysis of the loans in the database to predict Customer's payment behavior. Our Analysis consists of the following two parts :

- A segmentation model is carried out to determine different clusters of debtors and identify distinctive characteristics of each one of them.
- Develop a prediction algorithm that allows to determine the probability of payment of each loan.

Project Road Map

- Data Cleaning & Feature Engineering
- Exploratory Data Analysis
- Modelling and Evaluation
    - Classification
    - Cluster Analysis
- Conclusion



The Dataset

The Lending Club dataset used in this project has been taken from Kaggle:
[(https://www.kaggle.com/wordsforthewise/lending-club)](https://www.kaggle.com/wordsforthewise/lending-club)

The data is separated into 2 different files:

- Accepted loans (This is being used)
- Rejected loans

There are about 151 features of every loan of the dataset.

- Date range: January 2007 - December 2018 (11 years)
- Total of 2260701 rows of data

Data Cleaning & Feature Engineering

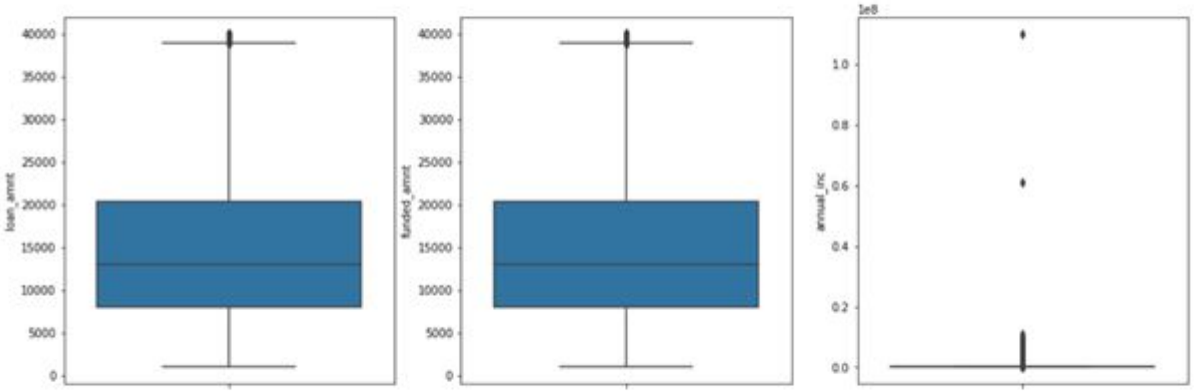
Data Cleaning

LendingClub

- In its original state, the data contains **2260701** observations and **151** variables. Data had the following drawbacks-
  - High number of observations
  - High dimensionality in the data.
- On observing the data, it was concluded that some variables are better described in the dictionary of information, so a match was done with this table to get better description of the variables.
- This helped to eliminate the variables that were not relevant as only the variables that were in both, the data and the dictionary, were only picked.
- Furthermore, the format of some variables containing dates was fixed and the ‘emp_length’ variable was transformed to a numeric variable.

Data Cleaning

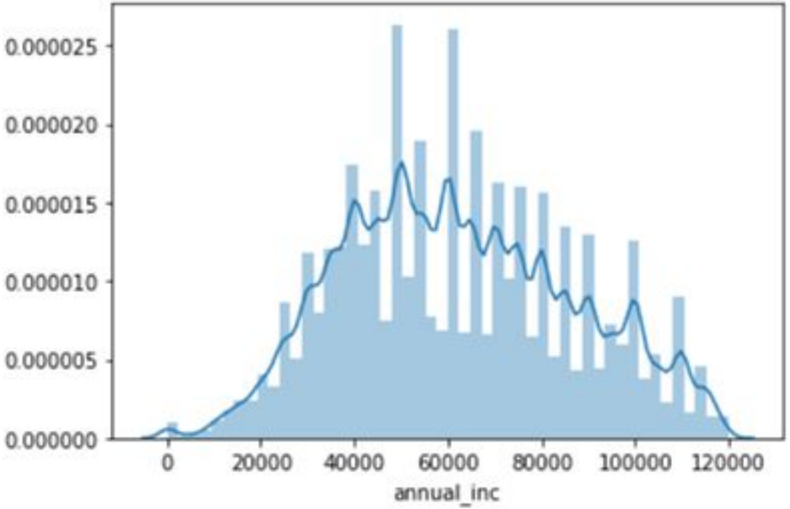
- Some basic boxplots were made to see the distribution of data points in the categories of loan amount, funded amount and annual income which can be seen below.

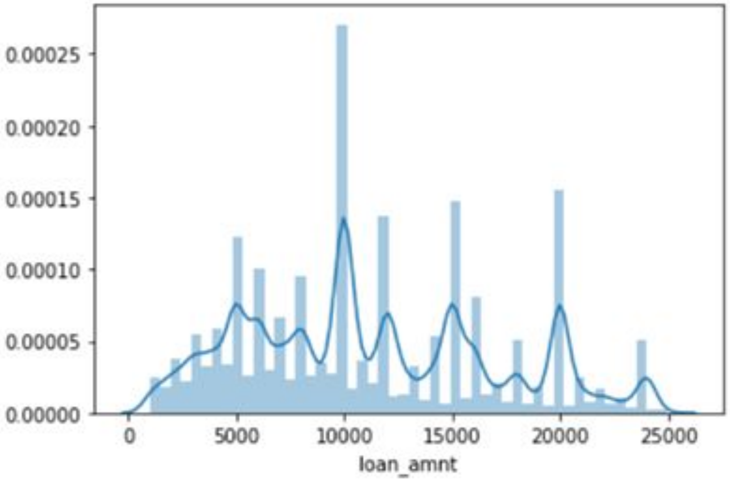


Data Cleaning

LendingClub

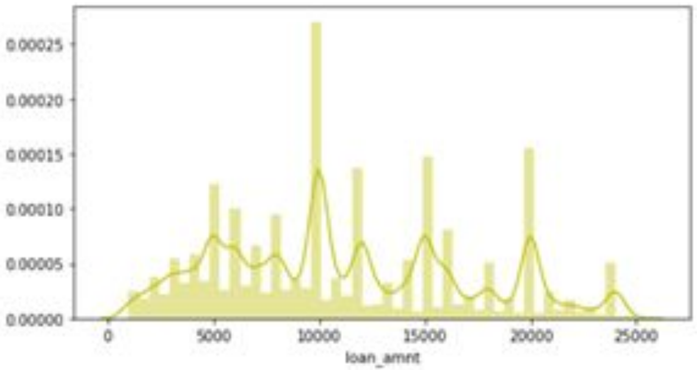
- Here we can see that most of the loan amounts and funded amounts fall in the range of 8000-20000, with some outliers.
- After removing the outlier frequency plots for loan and funded amounts were created which can be seen below.

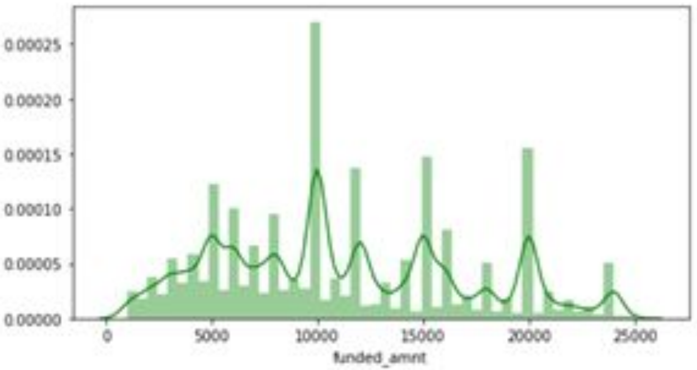




Data Cleaning

- After Outlier Treatment, we are still left with 72 % of data and we have sufficient information to proceed with Univariate Analysis.
- These variables are similarly distributed, which shows that there is an adequate balance between loan and funding.





Feature Engineering

LendingClub

Some of the data nuances were handled as follows:

- We fixed the format of the variables containing dates.
- We transformed the ‘emp_length’ variable for it to be numeric
- The NA values were handled in the following manner:
    - 'emp_title' and 'verification_status_joint' variables were filled with ‘ ’.
    - 'bc_open_to_buy', 'mo_sin_old_il_acct', 'mths_since_last_delinq', 'mths_since_last_major_derog', 'mths_since_last_record', 'mths_since_rcnt_il', 'mths_since_recent_bc', 'mths_since_recent_bc_dlq', 'mths_since_recent_inq', 'mths_since_recent_revol_delinq', 'pct_tl_nvr_dlq','sec_app_mths_since_last_major_derog' were filled with the max value of each column.
    - Rest of the columns were filled with the minimum value of each column.

In the end, the final dataset was left with **938821** observations and **102** variables.

The Final Dataset

Initial Dataset

- **2260701** observations and **151** variables

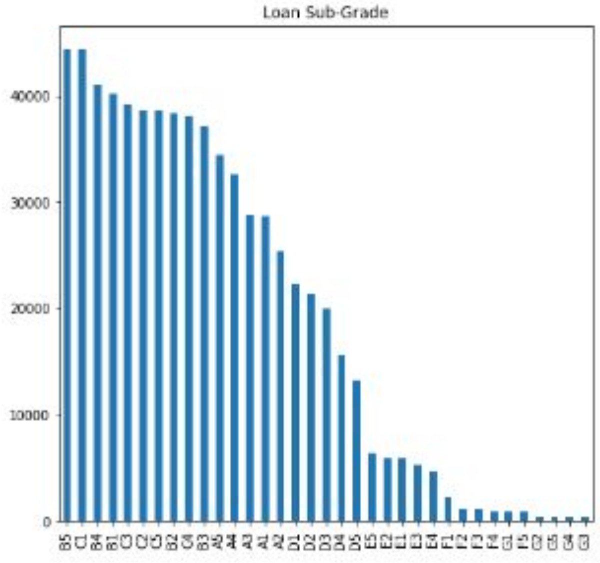
Final Dataset

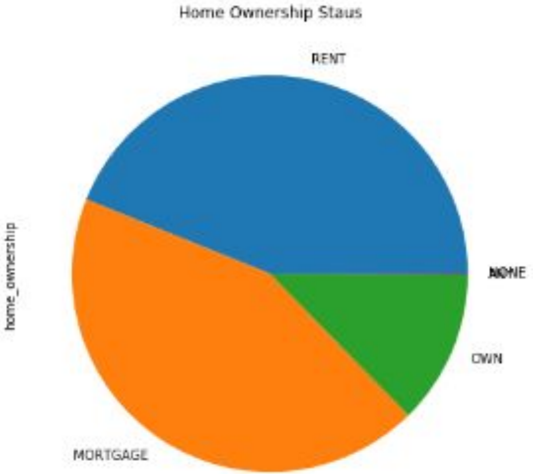
- **938821** observations and **102** variables

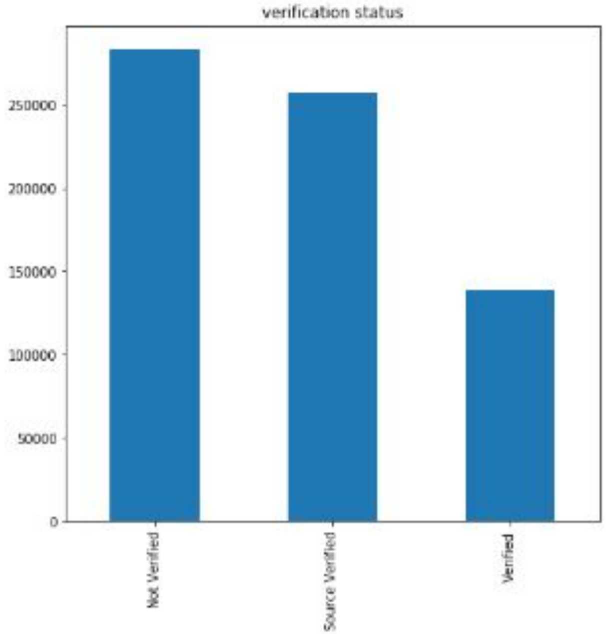
Exploratory Data Analysis

Loan Characteristics

- People who are taking loans have Home Ownership as Rent or in Mortgage.
- Most of loan applications do not have their income source verified, this is worth looking into as it might lead to defaulter loan

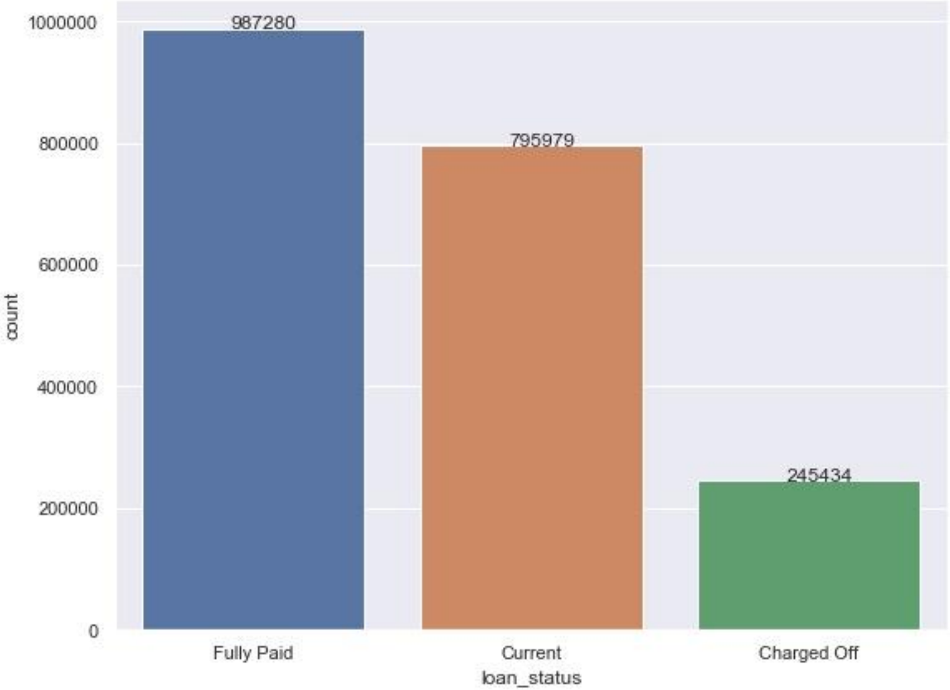






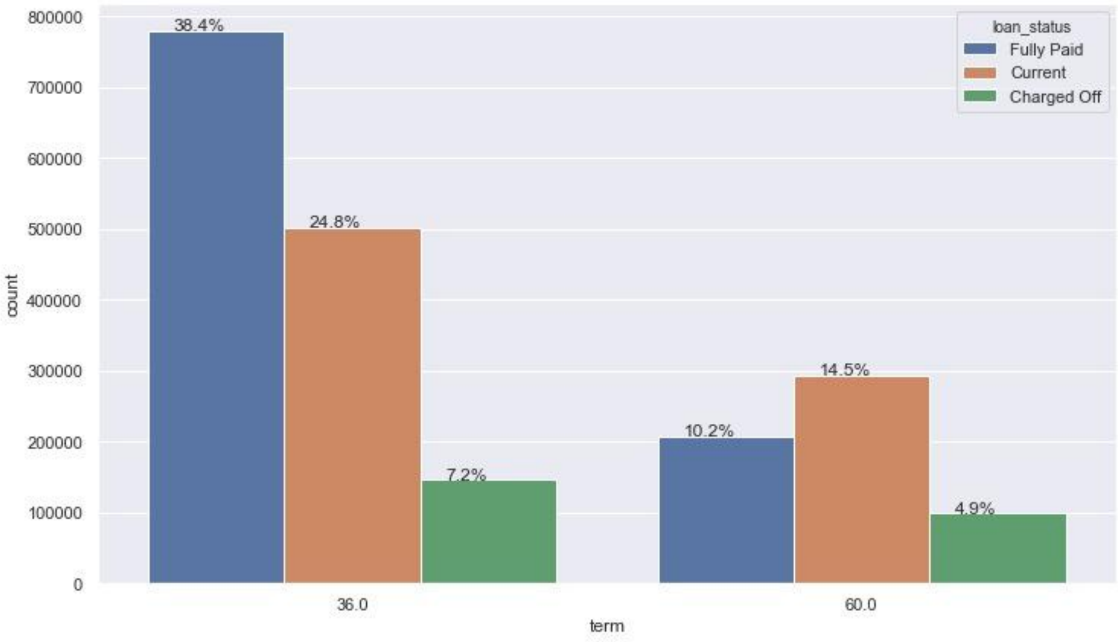
Loan Status

- We kept only 3 important loan statuses out of 7 present in the dataset, these are most useful
- We will focus on the Current and Charged off loans



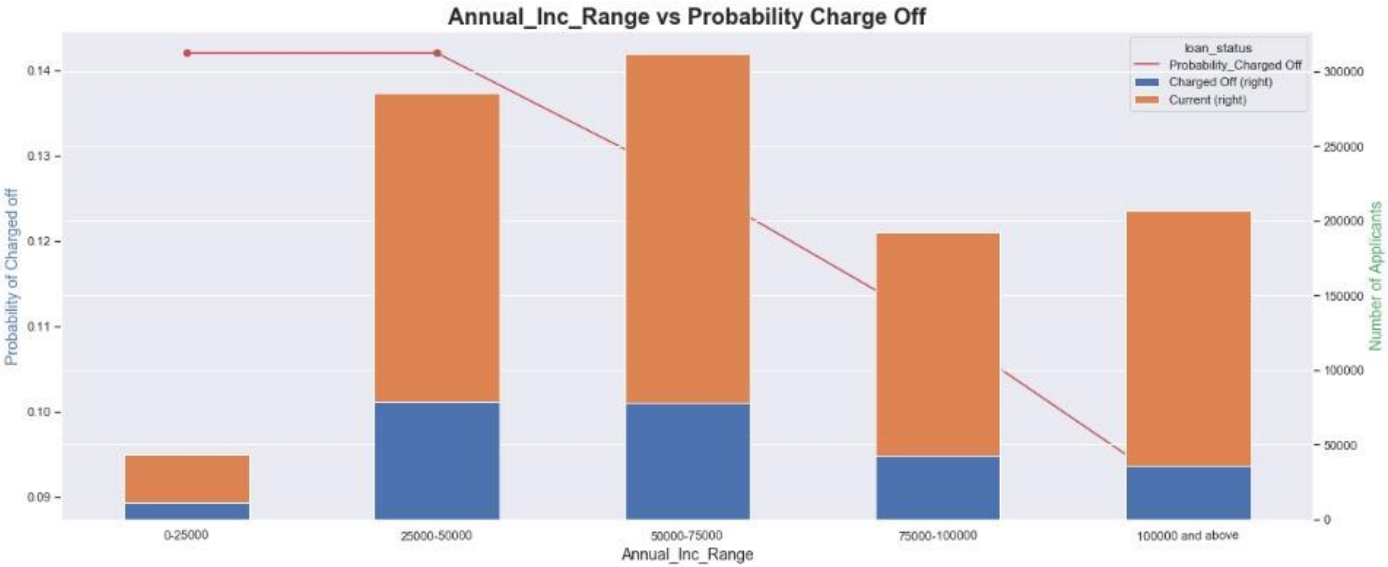
Loan Status Vs Loan Term

- There are only 2 loan terms - 36 months and 60 months
- Smaller term loans are more likely to be charged off compared to longer term loans but majority loans are short term



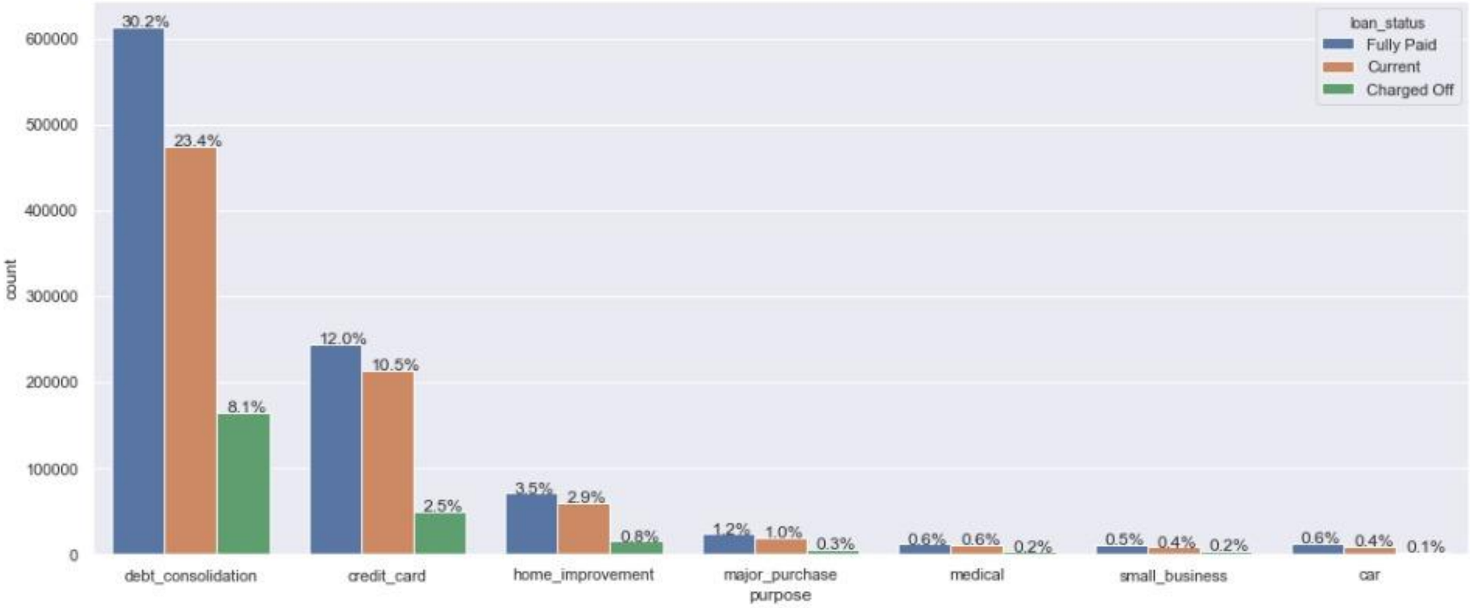
Annual Income Vs Probability Charge Off

- Annual income should play a huge role in determining loan charge off probability
- With the income increases the probability of charge off decreases drastically
- This can be an important feature for the model



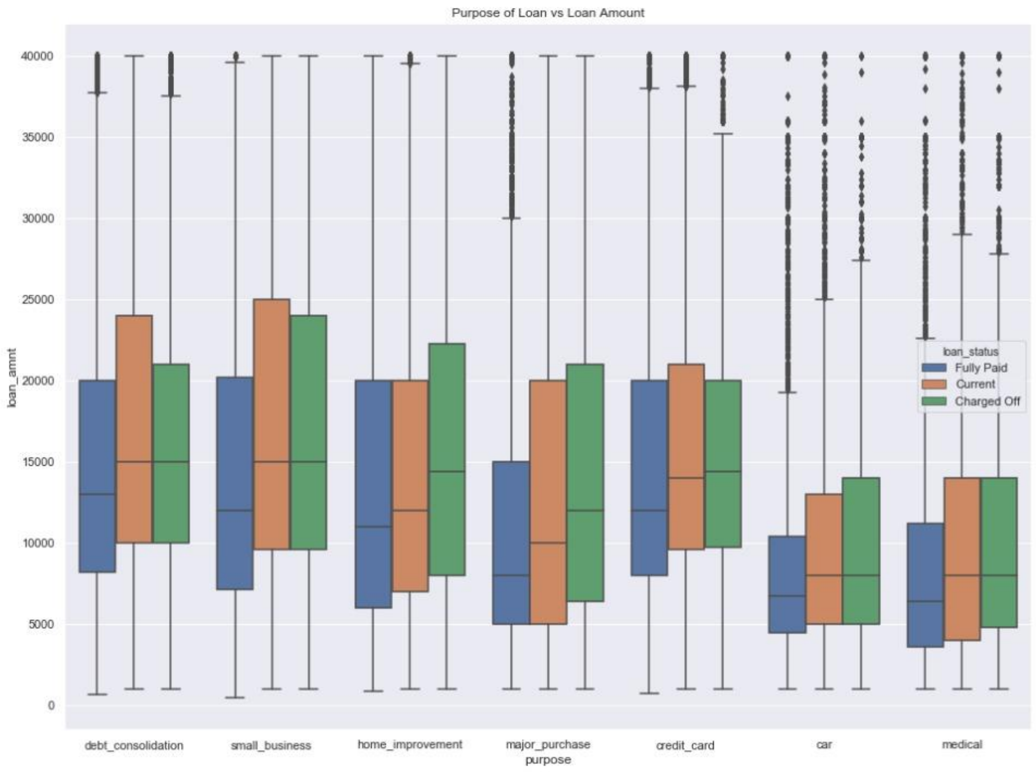
Purpose of Loan

- Most of the loans are taken for debt consolidation, credit card bills and home improvement and the charge off is also high for these loans



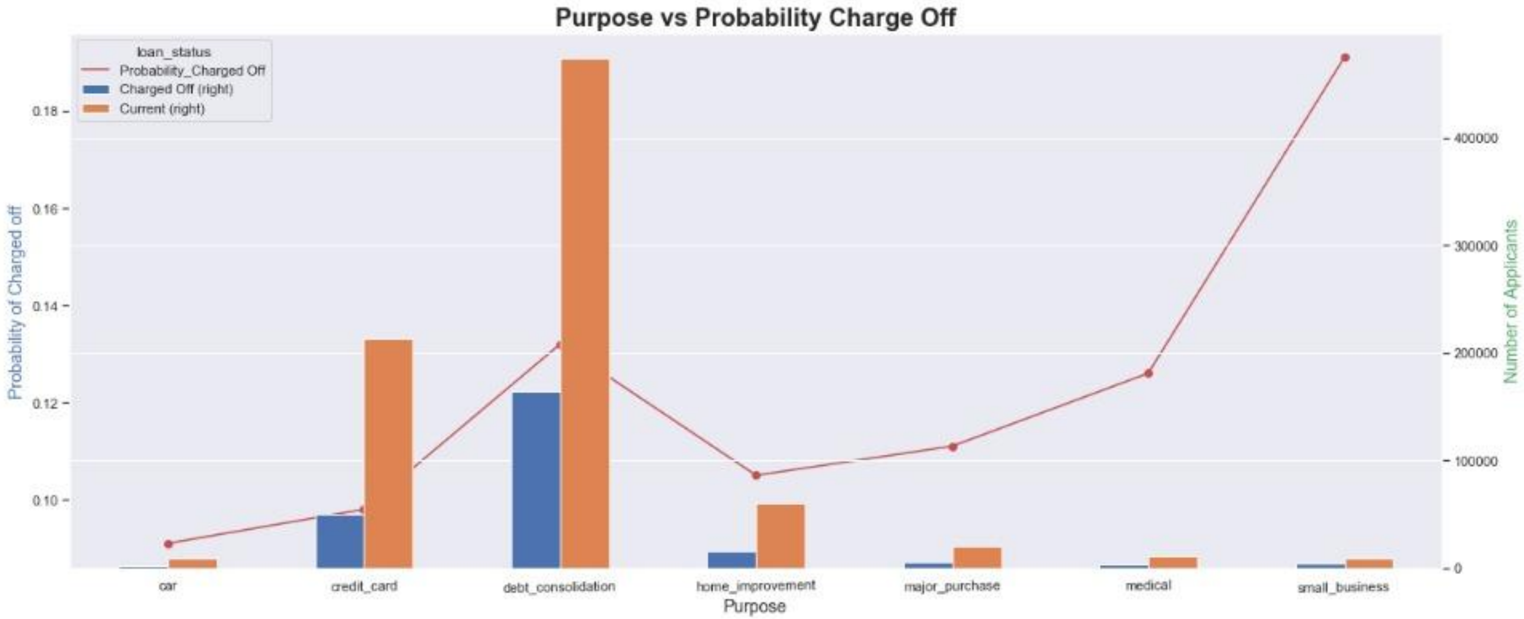
Purpose of Loan Vs Loan Amount

- For almost every loan purpose, the median loan amount for charged off loans is higher than the fully paid and current loans.
- Considering monitoring the loan amount would help reduce charge off probability



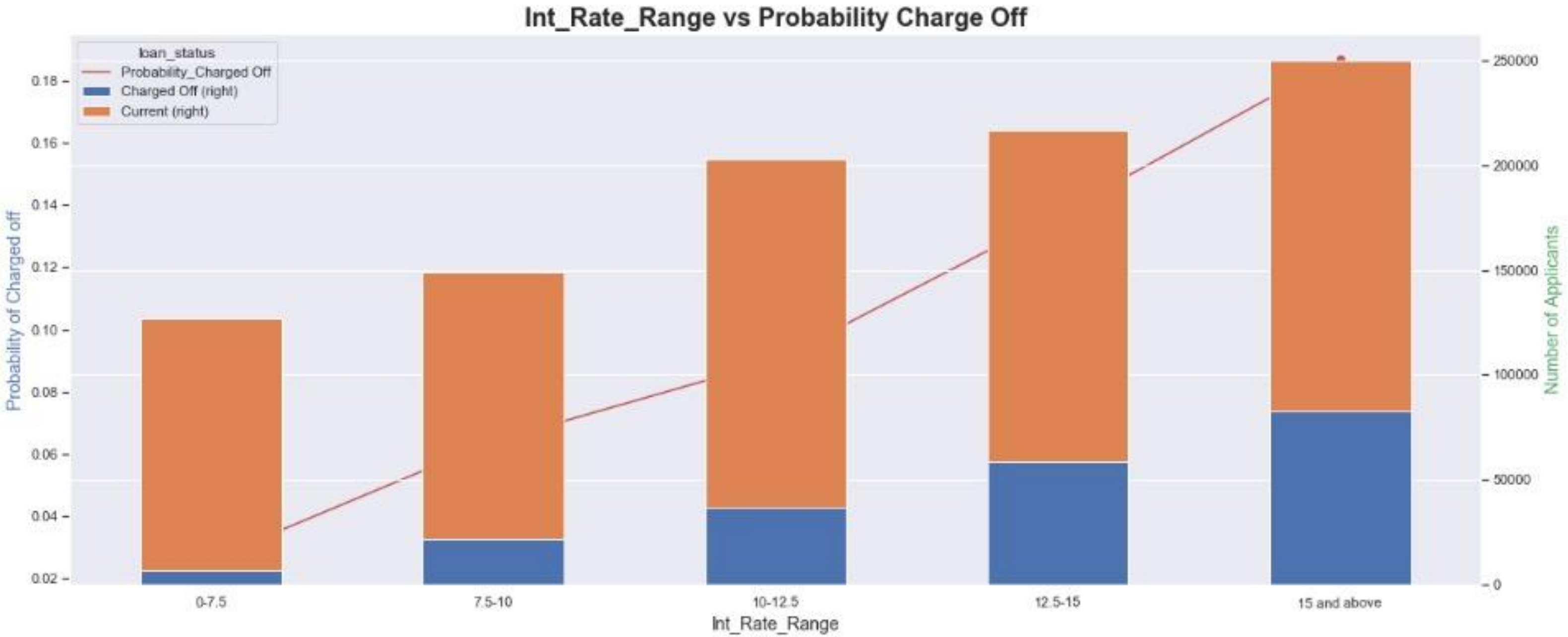
Purpose of Loan Vs Probability Charge Off

- Probability of charge off is really high for small business and debt consolidation loans.
- These types of loan should be monitored carefully



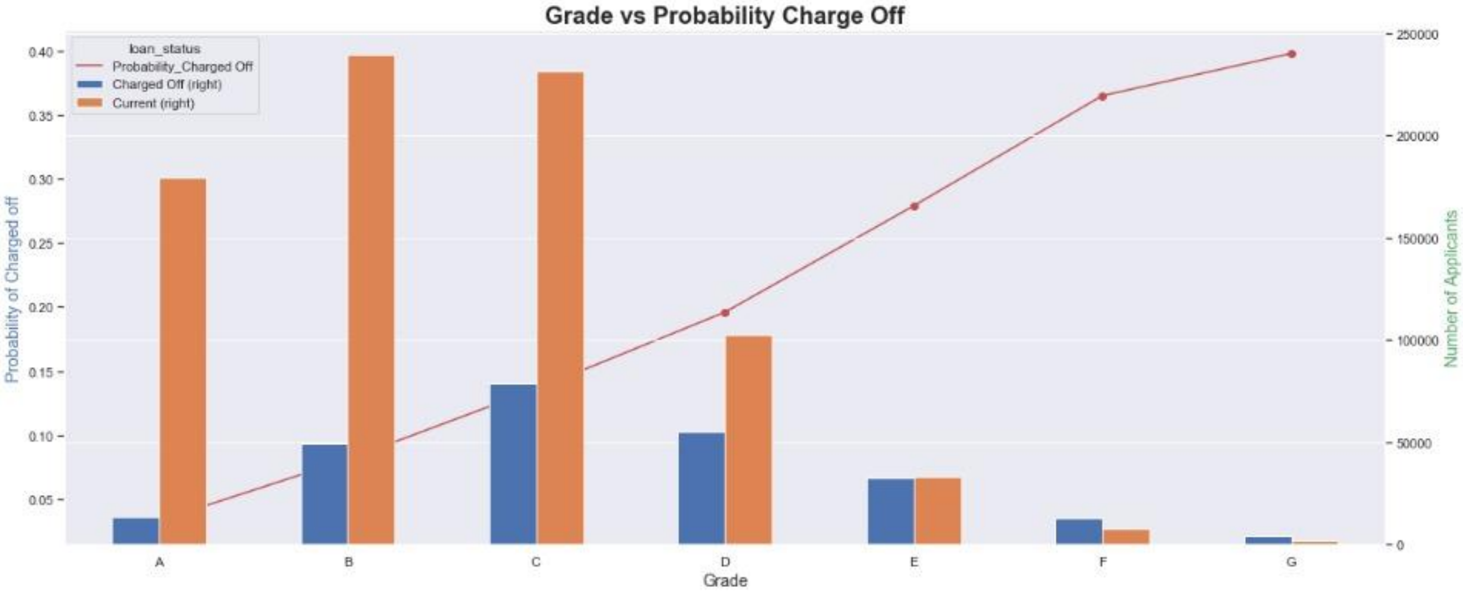
Interest Rate Range Vs Probability Charge Off

- Interest rate definitely affects the charge off loan.
- Loans with higher risk have high interest rate and in a way leads to charge off



Loan Grade Vs Probability Charge Off

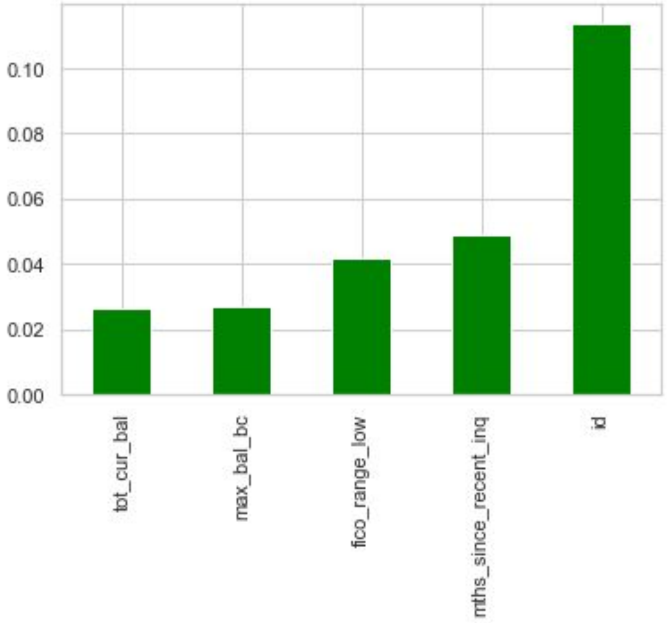
- Loan grades are decided by the risk of each loan hence the probability of charge off increases with the grade
- Due to this reason, risky grades have low approval count

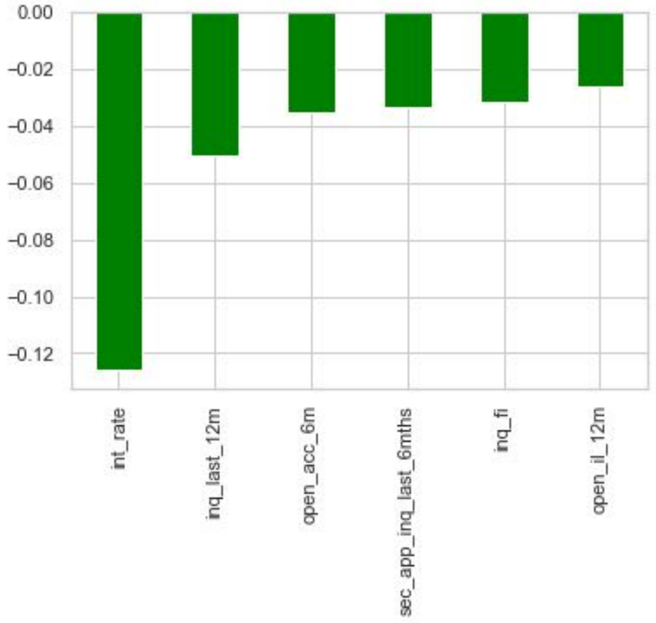


Highest Correlations with the outcome variable

LendingClub

- We computed the different feature correlations and saw which of them have the highest positive and negative correlations with the outcome variable.





Modelling

Models Used

LendingClub

- Classification Model
  - Naive Bayes Classifier
  - Random Forest
  - Logistic Regression
  - Neural Network

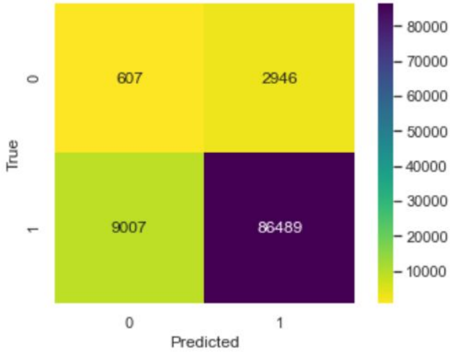
- Clustering Model
  - Kmeans

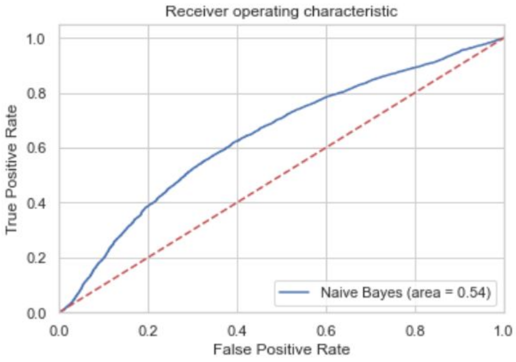
Classification - Naive Bayes Classifier

- 2018 info to achieve better running times
- Independence between each variable (Naive Bayes).

Model Accuracy: 0.879

	precision	recall	f1-score	support
0	0.06	0.17	0.09	3553
1	0.97	0.91	0.94	95496
accuracy			0.88	99049
macro avg	0.52	0.54	0.51	99049
weighted avg	0.93	0.88	0.91	99049





Feature Importance - Naive Bayes Classifier

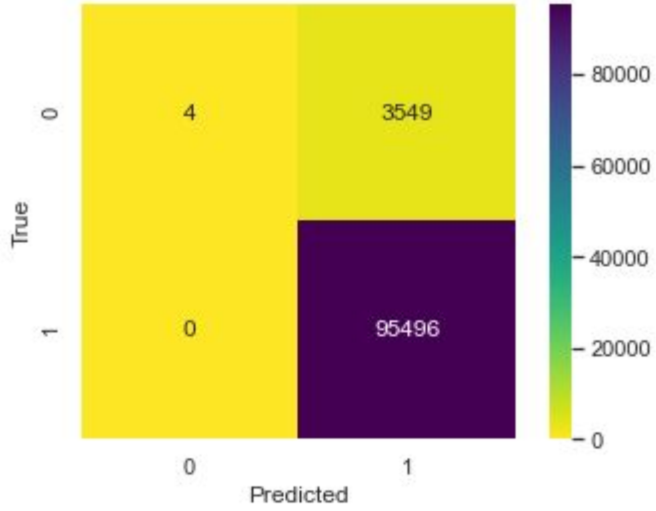
- We computed the feature importance of the Naive Bayes Classifier with the use of eli5 package in Python.

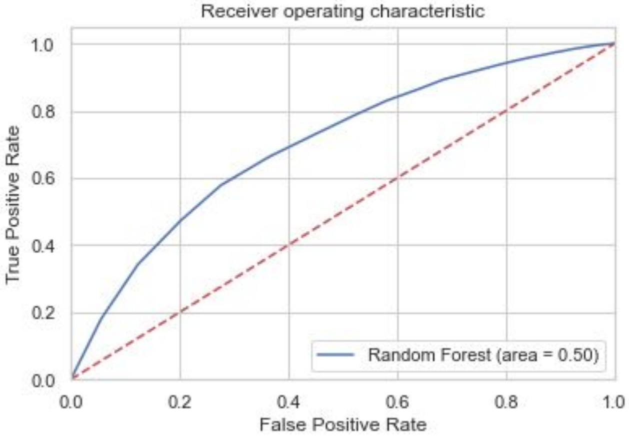
Weight	Feature
0.0023 ± 0.0005	sec_app_mths_since_last_major_derog
0.0012 ± 0.0003	bc_open_to_buy
0.0006 ± 0.0001	annual_inc_joint
0.0005 ± 0.0002	dti
0.0005 ± 0.0004	loan_amnt
0.0004 ± 0.0001	earliest_cr_line
0.0001 ± 0.0002	mths_since_recent_bc_dlq
0.0001 ± 0.0001	tot_coll_amt
0.0000 ± 0.0001	revol_util
0.0000 ± 0.0000	total_cu_tl
0.0000 ± 0.0000	grade_B
0.0000 ± 0.0002	emp_length
0.0000 ± 0.0001	delinq_amnt
0.0000 ± 0.0000	num_tl_90g_dpd_24m
0.0000 ± 0.0000	home_ownership_RENT
0 ± 0.0000	tax_liens
0 ± 0.0000	collections_12_mths_ex_med
0 ± 0.0000	grade_G
0 ± 0.0000	acc_now_delinq
0 ± 0.0000	grade_F
	… 36 more …

Classification - Random Forest

- 2018 info to achieve better running times
- Model Accuracy: 0.964

	precision	recall	f1-score	support
0	1.00	0.00	0.00	3553
1	0.96	1.00	0.98	95496
accuracy			0.96	99049
macro avg	0.98	0.50	0.49	99049
weighted avg	0.97	0.96	0.95	99049

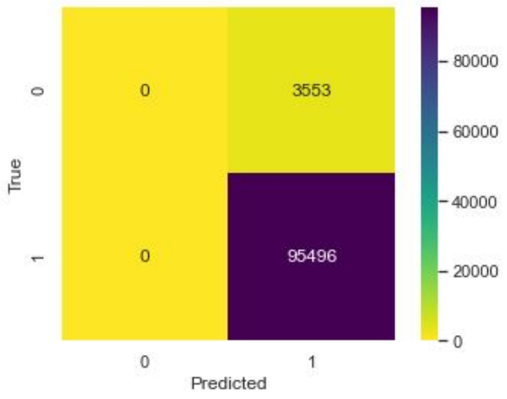


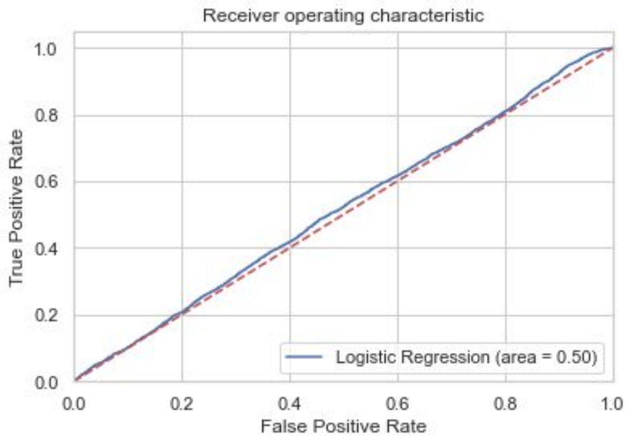


Classification - Logistic Regression

- Biggest running time of the models.
  - Picked only the most important features of the previous model to train the algorithm.
- Model Accuracy: 0.96

	precision	recall	f1-score	support
0	0.00	0.00	0.00	3553
1	0.96	1.00	0.98	95496
accuracy			0.96	99049
macro avg	0.48	0.50	0.49	99049
weighted avg	0.93	0.96	0.95	99049





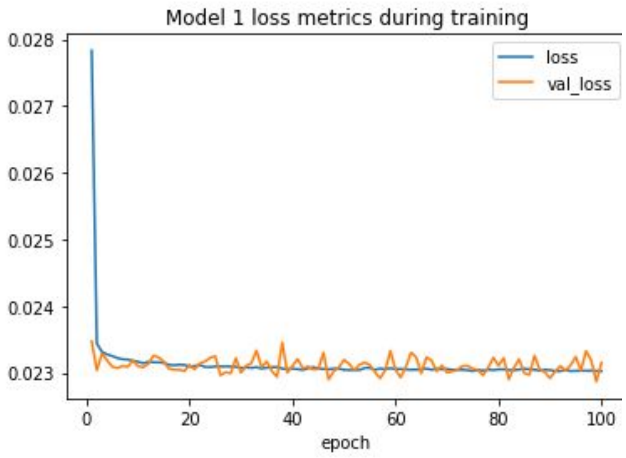
Classification - Neural Network Model

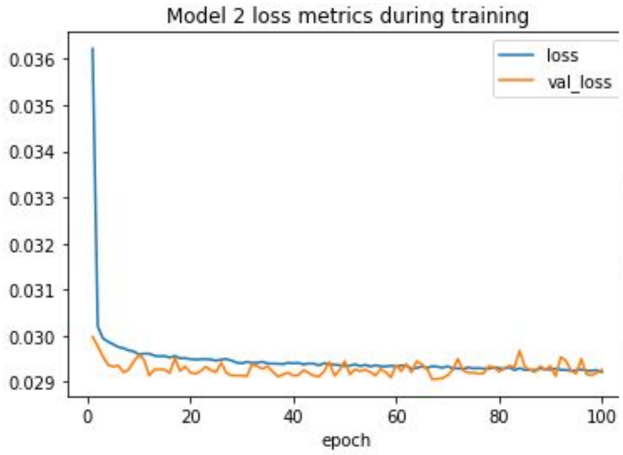
- 3 models with different dataset
- Model parameters:
  - Activation function - ReLu
  - Optimizer - Adam
  - Loss function - mean_squared_logarithmic_error

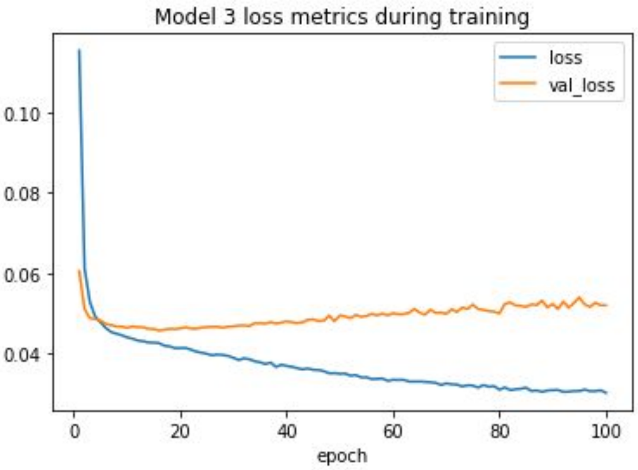
Layer (type)	Output Shape	Param #
dense_8 (Dense)	(None, 64)	5376
dropout_6 (Dropout)	(None, 64)	0
dense_9 (Dense)	(None, 32)	2080
dropout_7 (Dropout)	(None, 32)	0
dense_10 (Dense)	(None, 16)	528
dropout_8 (Dropout)	(None, 16)	0
dense_11 (Dense)	(None, 1)	17

Total params: 8,001
Trainable params: 8,001
Non-trainable params: 0

Classification - Neural Network Model Validation







Classification - Neural Network Validation (Model 2)

```
Train Result:
=====================================================
Accuracy Score: 80.54%
_____________________________________________
Classification Report:  Precision Score: 80.54%
                        Recall Score: 100.00%
                        F1 score: 89.22%
_____________________________________________
Confusion Matrix:
 [[    30 172815]
 [     4 715287]]

Train Result:
=====================================================
Accuracy Score: 80.54%
_____________________________________________
Classification Report:  Precision Score: 80.54%
                        Recall Score: 100.00%
                        F1 score: 89.22%
_____________________________________________
Confusion Matrix:
 [[     9  43205]
 [     5 178816]]
```

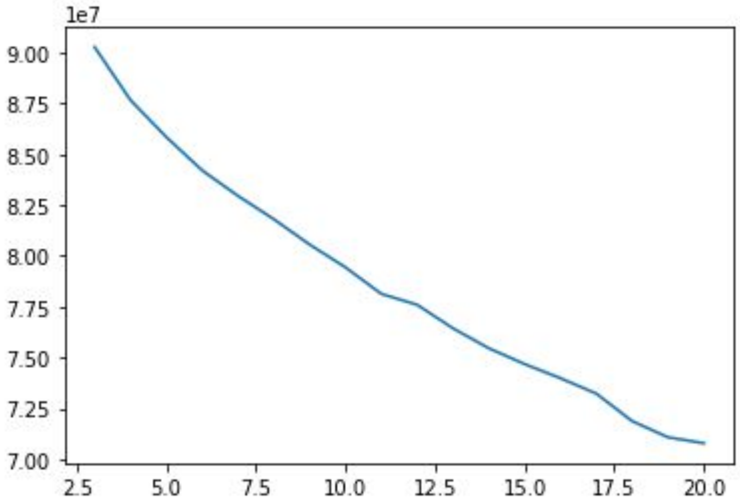
Classification Report:

	precision	recall	f1-score	support
0	0.64	0.00	0.00	43214
1	0.81	1.00	0.89	178821
accuracy			0.81	222035
macro avg	0.72	0.50	0.45	222035
weighted avg	0.77	0.81	0.72	222035

```
Confusion Matirx:
[[     9  43205]
 [     5 178816]]
```

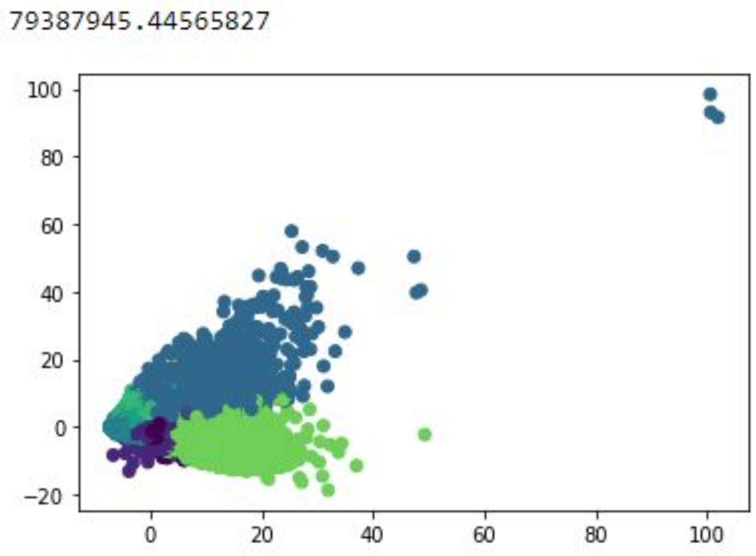
Clustering - KMeans Model

- Finding applicant segmentation to help understand different customer behaviors.
- Used PCA & KMeans algorithm to find clusters in the data



Clustering - KMeans Validation

- Unfortunately, there are no evident clusters in the data
- The loss graph does not flatten till k = 20
- Here is the cluster distribution for k = 10



LendingClub

Conclusion

Conclusion

- Some of the EDA is very helpful understanding the data.
- Precision is the best model performance metrics as we need to minimize False Positives.
- Machine learning models have much lower accuracy (& precision) compared to Deep Learning model.
- A neural network can be a good model to find the probability of charge off .
- Further analysis using CNN or RNN can help improve model accuracy.
- Future analysis to improve precision can include extensive feature extraction.
- Finding customer segmentation based on customer background can help label new applicants faster and decide loan grades.

Thank you!!