Natalia Katchoura

LendingClub
FINAL REPORT

Building a model that can predict whether or not a person will likely pay back a loan.

Problem Statement

The major problem for any financial companies is to estimate the risks associated with money loans, both for the borrower and the lender.

In my project I have built models to predict whether or not a person will likely pay back a loan.

To build my prediction model I was using LendingClub historical data of approved loans that can be obtained from Kaggle.
I examined 2.27M approved loans to analyse which of 151 characteristics affect the most the loans approval, the loan rates, terms and conditions, loan classification by grade.
I also reviewed if the approved loans were paid off (fully paid and charged off)

In my model I did not review missed or late payments and late fees.

Data Wrangling

With 2+ millions records and 151 parameters to consider I needed to do serious size reduction.

At first, I dropped all the columns where missing values exceeded 80%. Then I dropped columns with identical values (when >95% of records have the same value) and columns with information that cannot be used for my analysis such as in application_type, disbursement_method , emp_title, url columns.

151 loan features reduced to 108 which was still a big number.

The purpose of this project is to predict whether a loan will be fully repaid, so my target feature was loan_status.

The initial dataset had 9 unique values considering that I decided not to review missed and late payments or current loans.

```
Fully Paid                                             1076751
Current                                                 878317
Charged Off                                             268559
Late (31-120 days)                                       21467
In Grace Period                                           8436
Late (16-30 days)                                         4349
Does not meet the credit policy. Status:Fully Paid        1988
Does not meet the credit policy. Status:Charged Off        761
Default                                                     40
NaN                                                         31
Name: loan_status, dtype: int64
```

That gave me thousands of records which are not relevant to my analysis, so we can remove all the data that are not 'Fully Paid' or 'Charged off' from the dataset. The number of loans reduced from 2,260,699 to 1,345,310 in total. I had 1,076,751 fully paid loans and 268,559 charged off loans to train and test my model.

After all the data cleaning and wrangling I still had 108 features left (87 - numerical and 21 categorical).

For the model to be stable enough,it is advised that we keep only one feature in the dataset if two features are highly correlated. I found indices of feature columns and I dropped highly correlated features, the columns with correlation greater or equal 0.9.

Then based on Lending Club data dictionary
https://resources.lendingclub.com/LCDataDictionary.xlsx
I dropped all irrelevant or unknown by the time of the loan features. So I reduced my 87 numerical features to 57. Then I remove features with low variance. I had 49 numerical and 21 categorical features.

I reviewed my categorical features. For the further data analysis I convert some of them to numeric (such as length of employment, year, loan term) I also review categorical features and drop all categorical features with many levels
Finally I had the dataset with 1,345,310 records and 60 columns.

Exploratory Data Analysis

Approved loans features analysis

Analysis of characteristics that could affect the most the loans approval, the loan terms and conditions, loan classification by grade

LendingClub prime borrower averages:

- Credit score: 705.
- Income: $84,647.
- Average loan size: $15,369.
- Interest rate range: 8.46% - 20.74%.
- Common uses: credit card refinancing, new large purchases, debt consolidation.

LendingClub near-prime borrower averages:

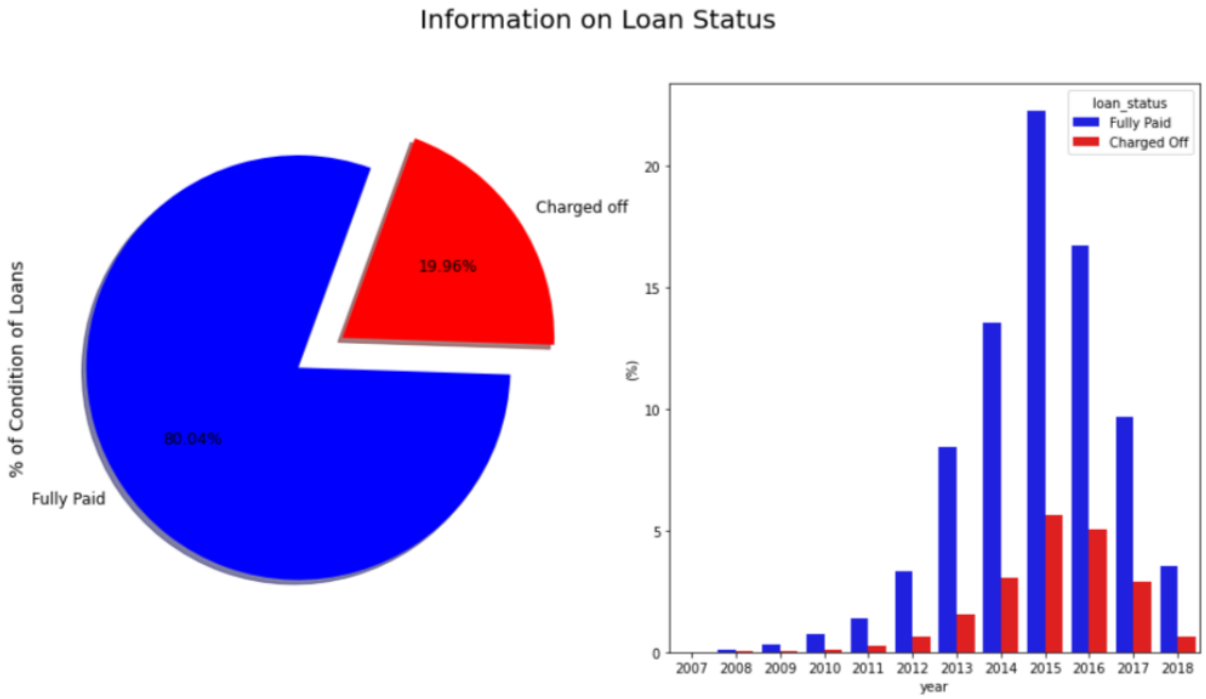
- Credit score: 640.
- Income: $69,082.
- Average loan size: $8,412.
- Interest rate range: 15.54% - 30.99%.
- Common uses: credit card refinancing, debt consolidation.

Overall statistics

	count	mean	std	min	25%	50%	75%	max
loan_amnt	2260668.0	15046.931228	9190.245488	500.00	8000.00	12900.00	20000.00	40000.00
funded_amnt	2260668.0	15041.664057	9188.413022	500.00	8000.00	12875.00	20000.00	40000.00
funded_amnt_inv	2260668.0	15023.437745	9192.331679	0.00	8000.00	12800.00	20000.00	40000.00
int_rate	2260668.0	13.092829	4.832138	5.31	9.49	12.62	15.99	30.99
installment	2260668.0	445.806823	267.173535	4.93	251.65	377.99	593.32	1719.83
...	...	...	...	...	...	...	...	...
tax_liens	2260563.0	0.046771	0.377534	0.00	0.00	0.00	0.00	85.00
tot_hi_cred_lim	2190392.0	178242.753744	181574.814655	0.00	50731.00	114298.50	257755.00	9999999.00
total_bal_ex_mort	2210638.0	51022.938462	49911.235666	0.00	20892.00	37864.00	64350.00	3408095.00
total_bc_limit	2210638.0	23193.768173	23006.558239	0.00	8300.00	16300.00	30300.00	1569000.00
total_il_high_credit_limit	2190392.0	43732.013476	45072.982191	0.00	15000.00	32696.00	58804.25	2118996.00

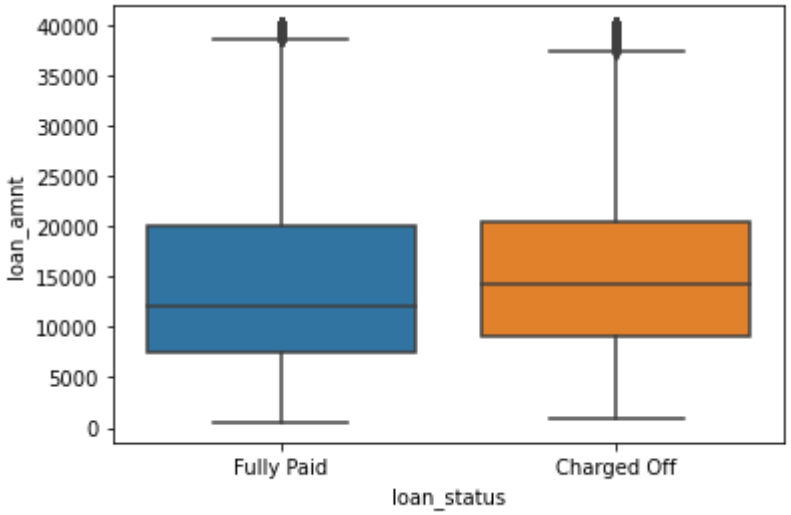
87 rows × 8 columns

Information on loans status



We know the number of prepaid and charged off loans (see above)
Let's plot the percentage of 'Fully Paid' vs 'Charged Off' loans and how loans were distributed by loan status over the years.
Over the years more hen 80% of all loans were fully repaid

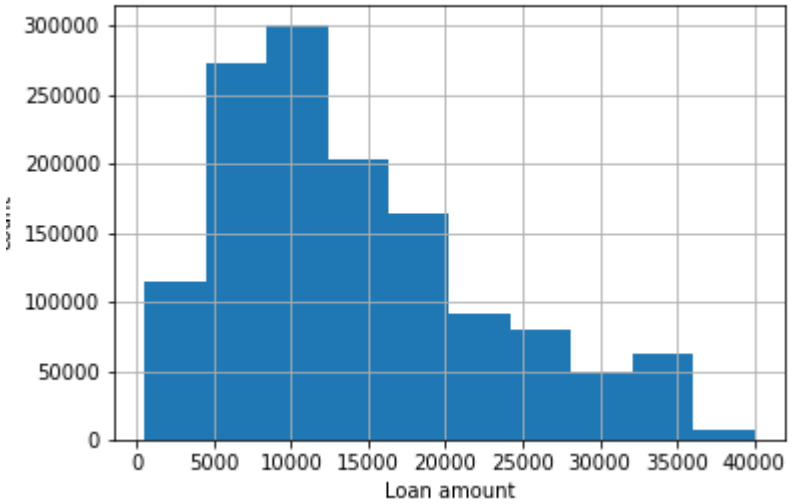
Loan Status vs. Loan amount



The box plot above shows that there is no significant difference in loan amounts for prepaid and charged off loans.

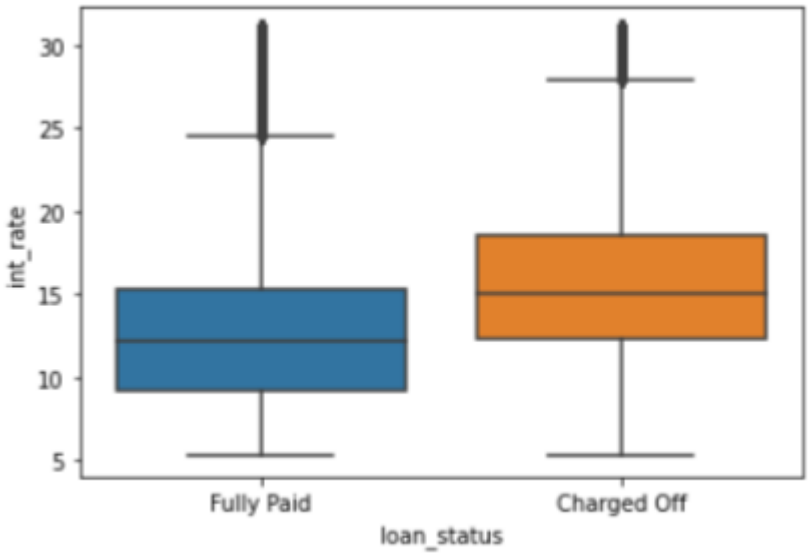
Loan amount distribution

The Lending Club loan amounts vary from 1K to 40K



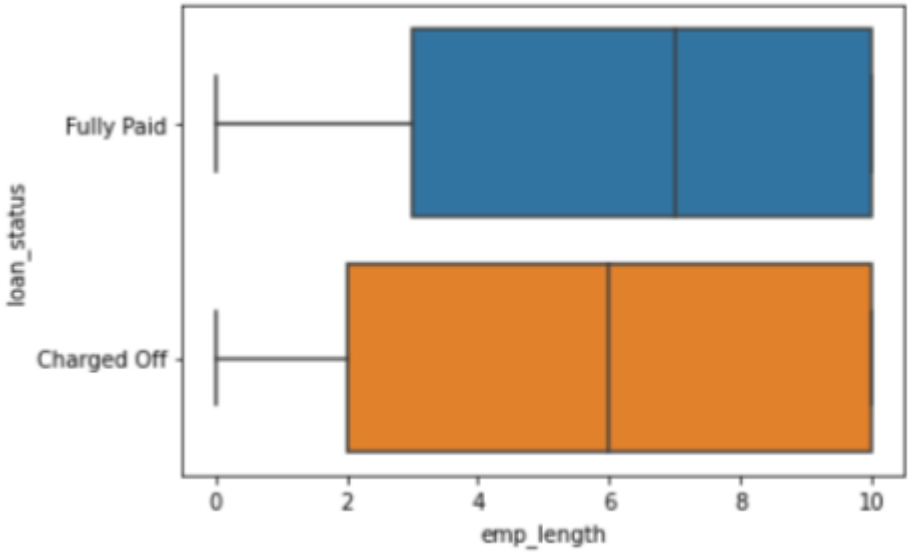
The loan amount distribution graph shows that most loans are between 5K to 20K range

Loan status vs. Interest rates



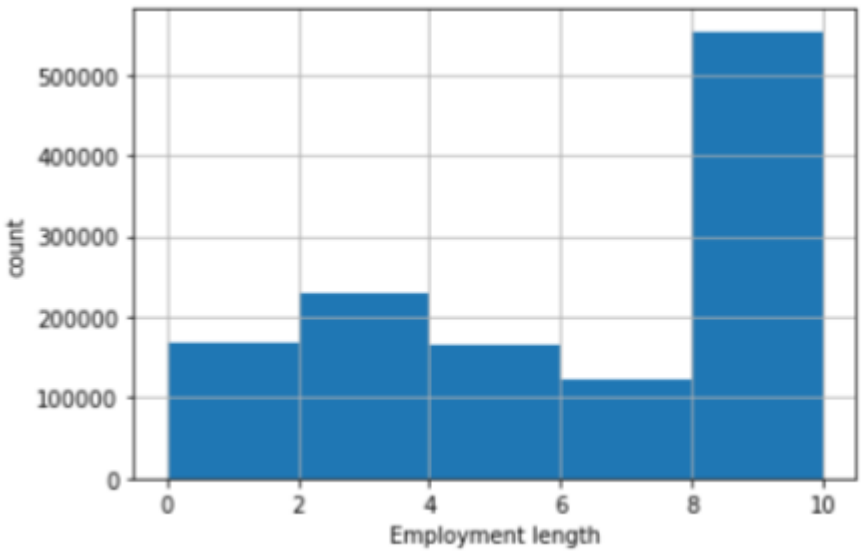
Our dataset contains loans with the interest rates as low as 7% and as high as 30% According to our box plot the loans with lower interest rate more likely to be fully paid

Loan status vs. Employment length



The boxplot displays that there is no significant difference in loan status vs employment length

Statistics of Employment length



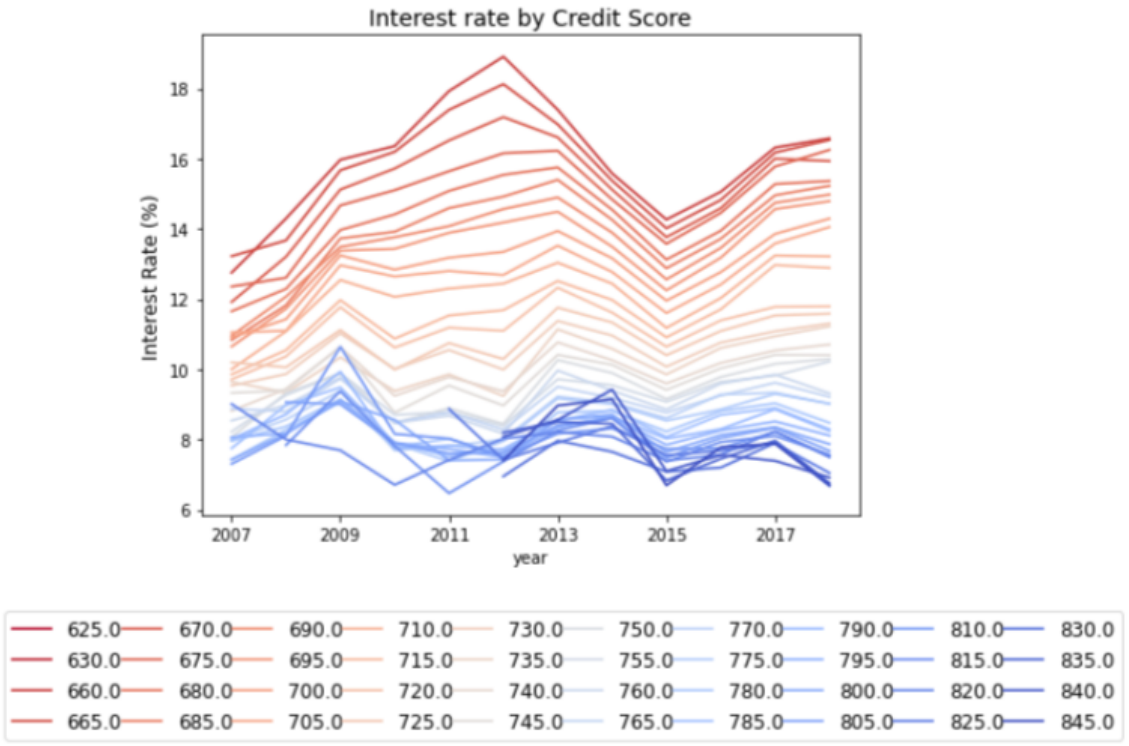
Most of the borrowers have more than 8 years of employment length. That explains the previous boxplot.

Dependencies between the loan_status and the grade

Loan grading is a classification system that involves assigning a quality score to a loan based on a borrower's credit history, quality of the collateral, and the likelihood of repayment of the principal and interest.
There are 6 types of grades from 'A' (highest) to 'G'(lowest).

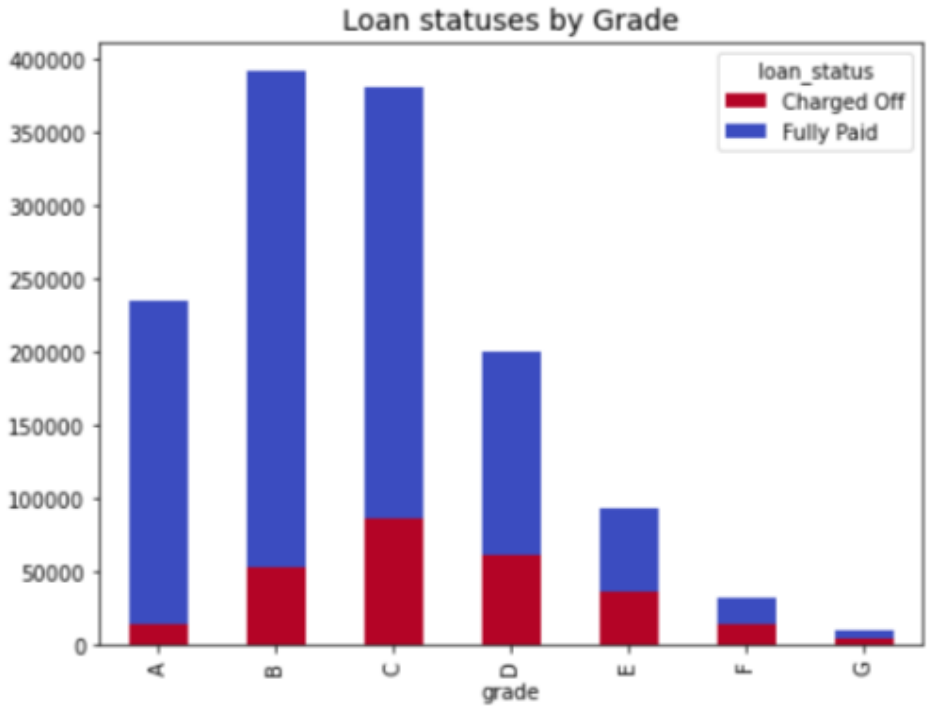
In our dataset the ratio of higher grades (less risky) loans is higher than lower grades (more risky) loans.

Average Interest rate by Credit Score



The expectations are the higher a credit score the lower the interest rate.The plot above shows that borrowers with the higher credit score have lower interest rate

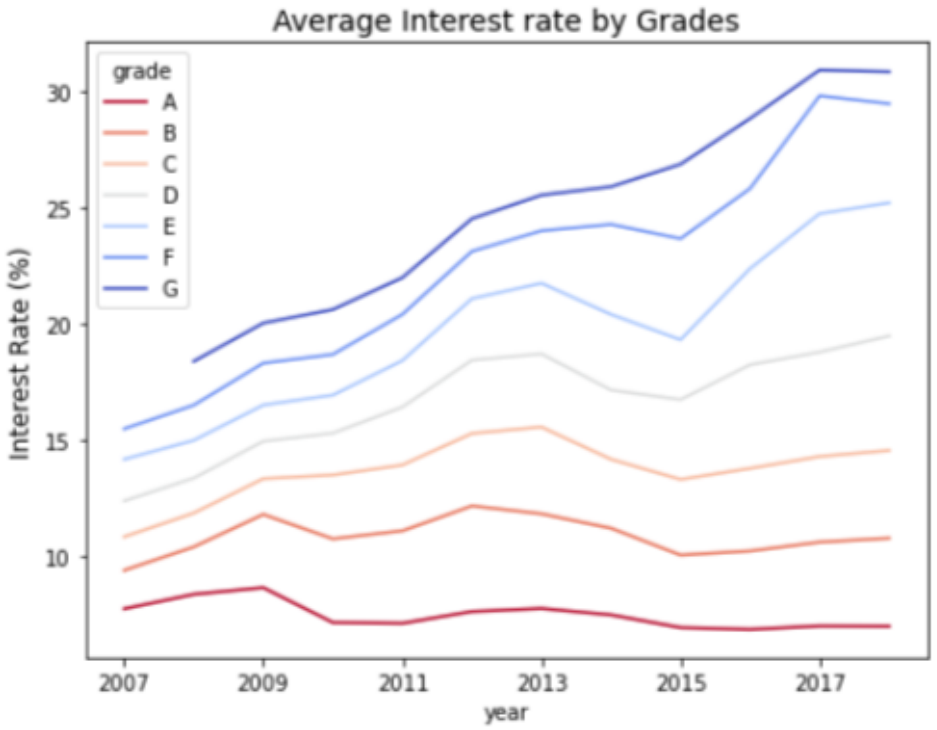
Loan statuses by Grade



The graph above shows as expected the higher grades proportionally have less charged off loans then the lower grades.

Average interest rate by Grade

We also expect that the interest rate will depend on the loan grade. The lower grade loans should have a higher interest rate.



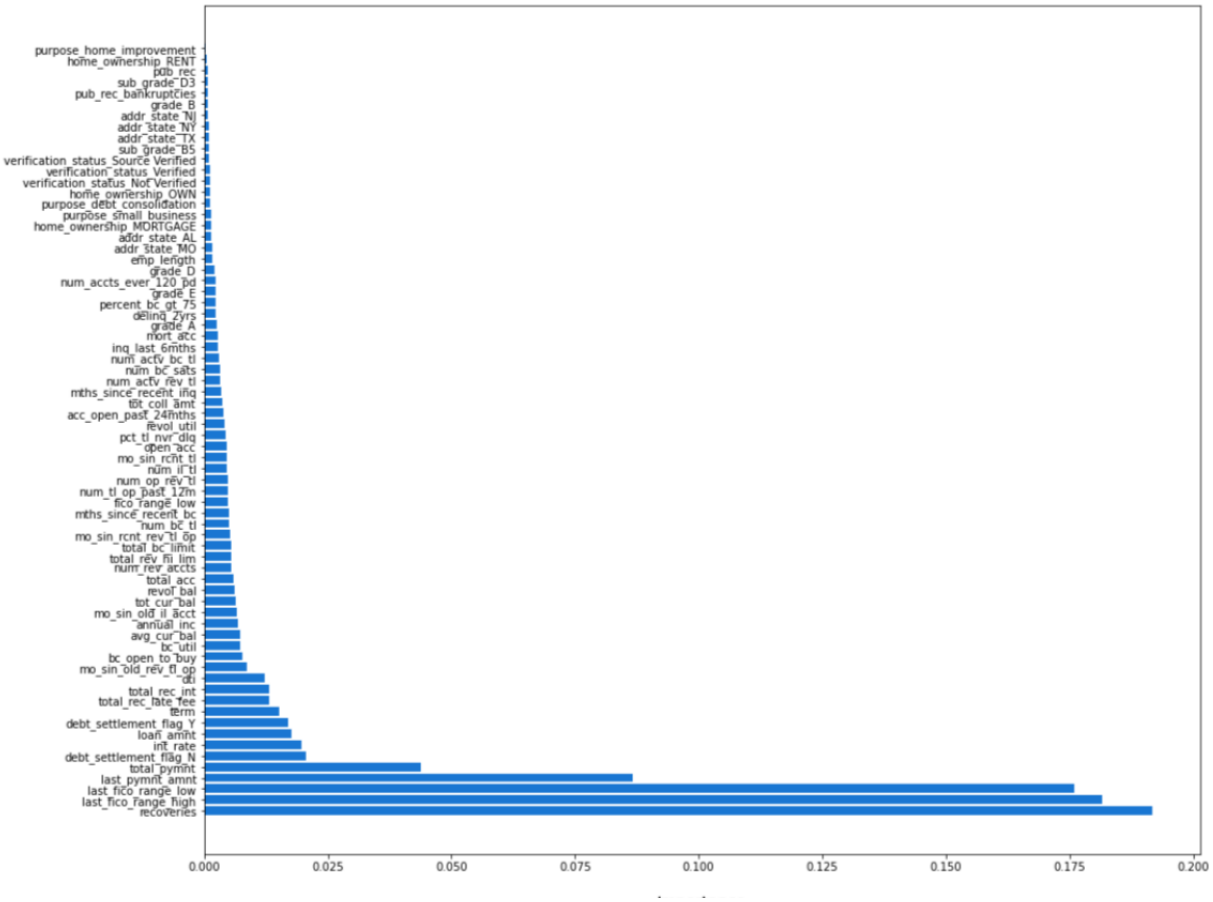
Comparing the loans grade 'A' and 'F' there is a bid difference in interest_rate mean A = 7.1 vs. mean F 24.9 max A = 9.6 vs max rate F = 30.75

Recursive feature selection

I used a Recursive feature elimination algorithm to select optimal features (Recursive feature elimination algorithm with Cross-Validation) and to remove the weakest features for 3 cases: the first 1,000 rows of our dataset, 10,000 first rows and 1,000 random selected rows.

Plot feature importances

RFECV - Feature Importances first 1,000 records



The result of Recursive feature elimination for the first 1K and 10K row dataset displayed the same result - 70 features, see the plot above.

The result of recursive feature elimination for randomly selected 1K records is very different. We ended up with 10 important features (recoveries, last_fisco_range_high,last_fisco_range_low, last_pymnt_amount,total_pymnt, debt_settlement_flag_N, int_rate,loan_amnt).

Result for the 1,000 randomly selected records - following 9 optimal features:

	loan_amnt	int_rate	total_pymnt	total_rec_int	recoveries	last_pymnt_amnt	last_fico_range_high	last_fico_range_low	debt_settlement_flag_N
0	7400.0	19.99	7469.850167	69.85	0.00	7486.29	704.0	700.0	1
1	18250.0	11.14	20012.516107	1762.52	0.00	13426.96	724.0	720.0	1
2	28800.0	27.34	8892.990000	1865.21	6324.64	885.28	584.0	580.0	1
3	6200.0	5.93	6781.057552	581.06	0.00	188.40	659.0	655.0	1
4	6500.0	15.31	6182.160000	1498.68	214.84	535.64	514.0	510.0	1

I had different number of optimal features from running first 1K records, 1K randomly selected records and 10K first records but the first most optimal features are the same:

- debt_settlement_flag_N

- debt_settlement_flag_Y
- total_pymnt
- last_pymnt_amnt
- recoveries
- revol_util
- total_rec_int
- total_rec_late_fee

Modeling (Classification models)

My goal was to create a model that predicts the value of a target variable by learning simple decision rules inferred from the optimal data features.
This is a classification problem. Here we have used the following classification models:

- Random Forest
- K-Nearest Neighbor (KNN)
- Naive Bayes
- Gradient Boost

And select the best performing model.

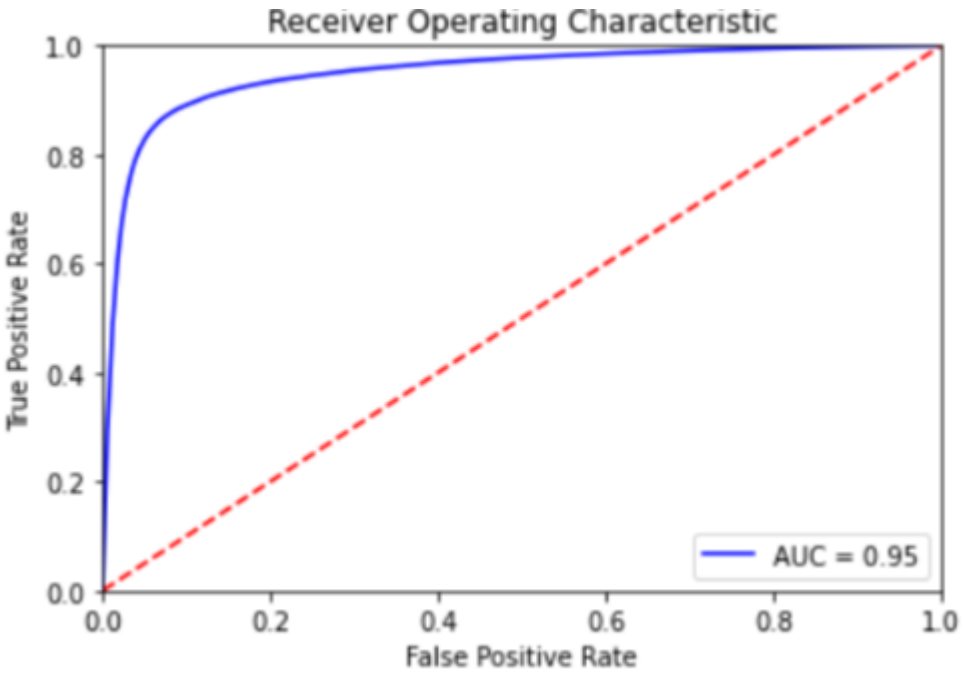
Classification algorithm

Random Forest

Classification report

	precision	recall	f1-score	support
0	0.69	0.76	0.73	52552
1	0.94	0.91	0.92	203350
accuracy			0.88	255902
macro avg	0.82	0.84	0.83	255902
weighted avg	0.89	0.88	0.88	255902

```
Accuracy: 0.8818883791451415
Balanced accuracy: 0.8378686856996
Precision score 0.9371468106273704
Recall score 0.7631679098797381
F1 score 0.8839555028220252
```



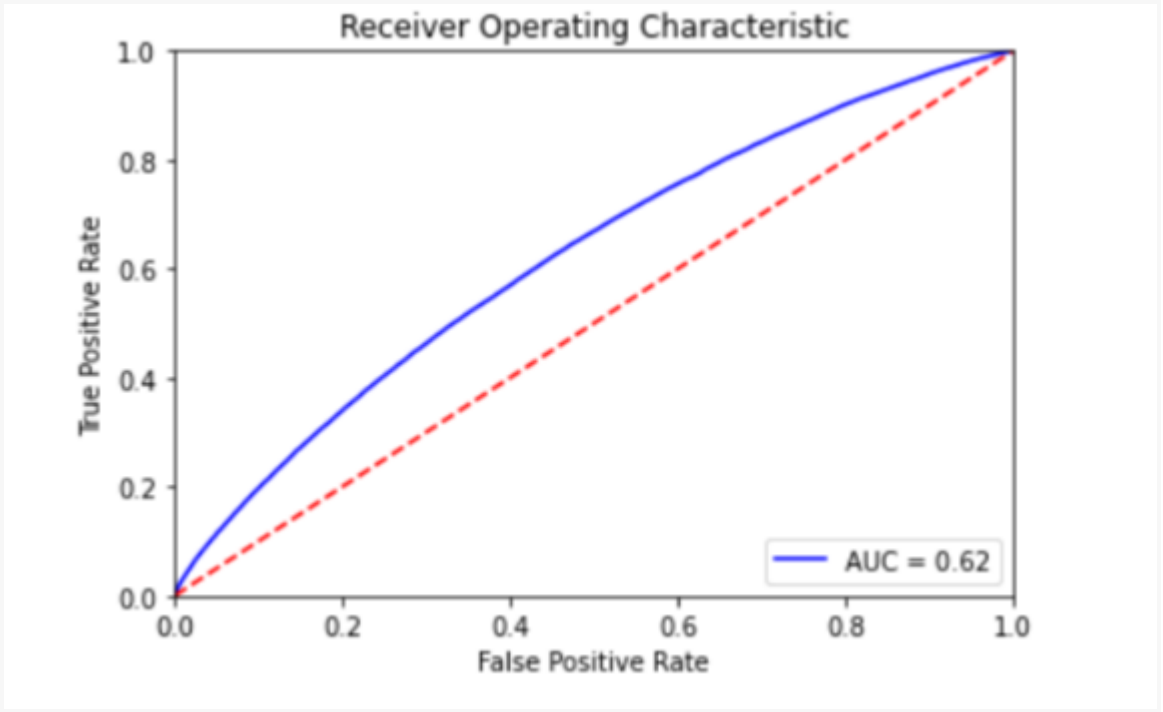
Compute the area under the Receiver Operating Characteristic Curve (ROC AUC) from prediction scores. - **0.8551925669134215**

K-nearest neighbor

Classification report

	precision	recall	f1-score	support
0	0.46	0.01	0.02	52811
1	0.79	1.00	0.88	203091
accuracy			0.79	255902
macro avg	0.63	0.50	0.45	255902
weighted avg	0.73	0.79	0.71	255902

```
Accuracy: 0.7933271330431181
Balanced accuracy: 0.5029069819033505
Precision score 0.7945838448896019
Recall score 0.00836946848194505
F1 score 0.7053821660739947
```



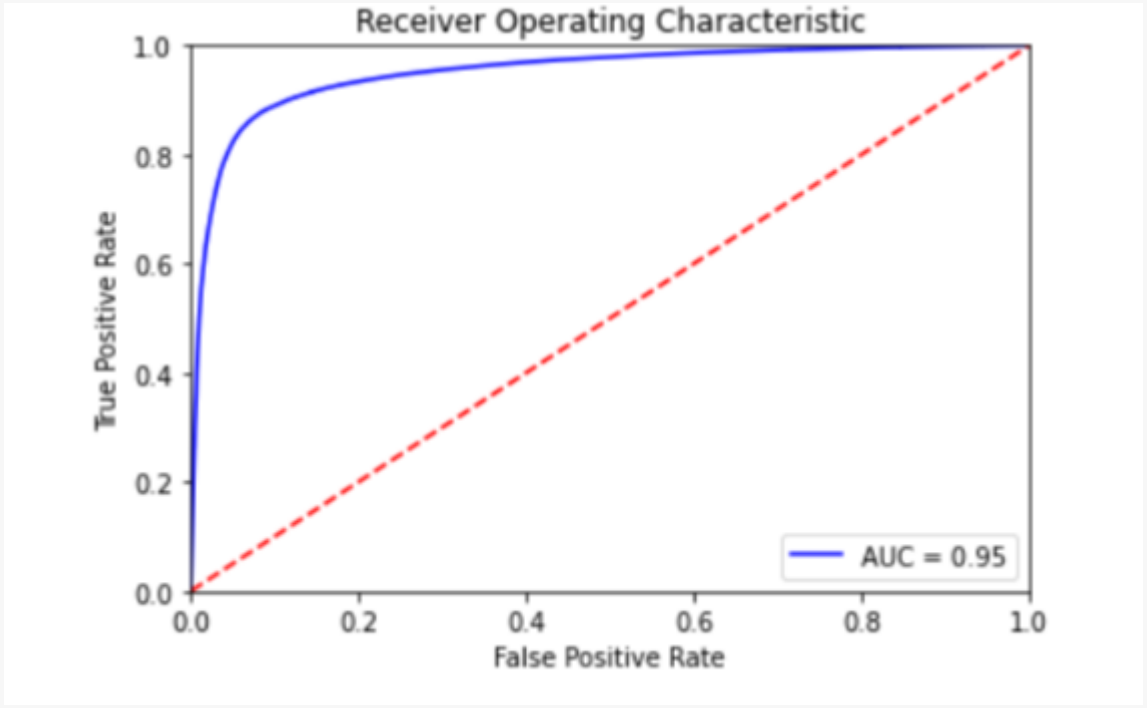
Compute the area under the Receiver Operating Characteristic Curve (ROC AUC) from prediction scores. - **0.5029069819033505**

Gradient Boosting

Classification report

	precision	recall	f1-score	support
0	0.77	0.78	0.77	52811
1	0.94	0.94	0.94	203091
accuracy			0.91	255902
macro avg	0.86	0.86	0.86	255902
weighted avg	0.91	0.91	0.91	255902

```
Accuracy: 0.9062101898382975
Balanced accuracy: 0.8585438401495943
Precision score 0.9419705435233263
Recall score 0.7773759254700725
F1 score 0.9063691265633111
```



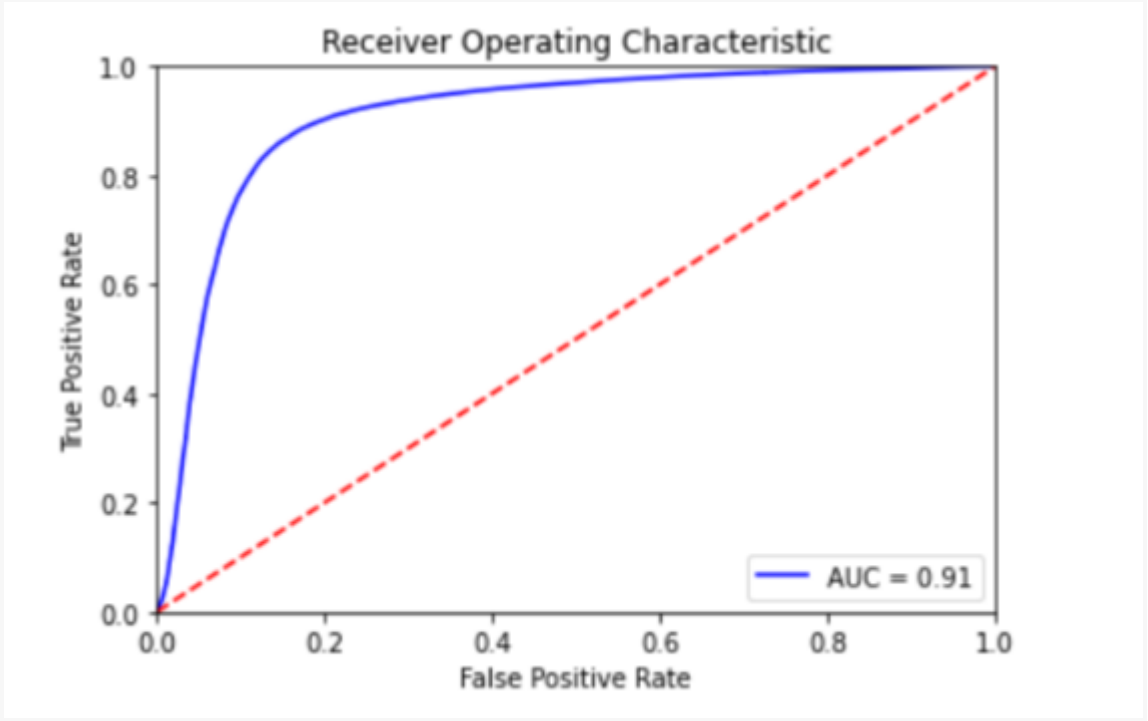
Compute the area under the Receiver Operating Characteristic Curve (ROC AUC) from prediction scores - **0.8585438401495943**

Naive Bayes

Classification report

	precision	recall	f1-score	support
0	0.56	0.88	0.68	52811
1	0.96	0.82	0.88	203091
accuracy			0.83	255902
macro avg	0.76	0.85	0.78	255902
weighted avg	0.88	0.83	0.84	255902

```
Accuracy: 0.8309001101984353
Balanced accuracy: 0.8493416491874122
Precision score 0.9634716438340274
Recall score 0.8807445418568102
F1 score 0.8430225459321076
```



Compute the area under the Receiver Operating Characteristic Curve (ROC AUC) from prediction scores.- **0.8493416491874121**

I have evaluated each model in terms of model accuracy, precision, recall and 'ROC-AUC' score for both the training and test data, and plotted them.

To select the best performing model I need to select the best metrics.

Comparison of the models

I had a binary classification problem. Our records belonging to two classes fully paid - 1 and charged off - 1 (YES or NO). On testing our model on 255902 samples, I got the following result.

I applied different ML algorithms to select the best performing models I started by choosing the right metric for evaluating machine learning ML models.

Metrics to evaluate ML models:

1.**Accuracy**
Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of the number of correct predictions to the total number of input samples. It works well only if there are an equal number of samples belonging to each. That is not our case:

812275 borrowers fully paid their loans and 211333 = charged off.

approximately 80% vs. 20%

In this case, even if you predict all samples as the most frequent class you would get a high accuracy rate, which does not make sense at all (because your model is not learning anything, and is just predicting everything as the top class)

2. **Precision**
Precision (#samples correctly predicted/#samples predicted) is a valid choice of evaluation metric when we want to be very sure of our prediction Being very precise means our model will leave a lot of credit defaulters untouched and hence lose money.

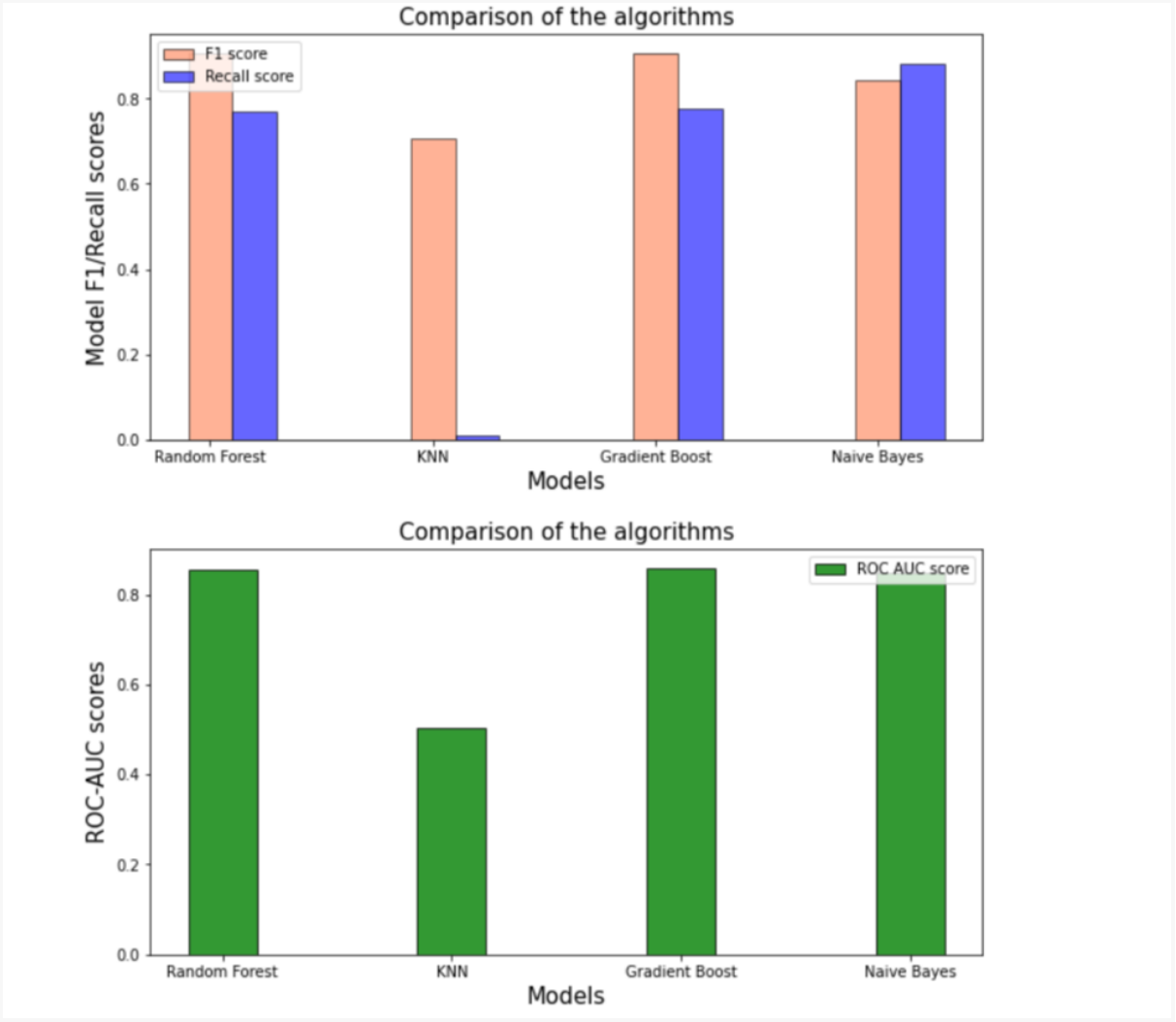
3. **Recall**
Recall is a valid choice of evaluation metric when we want to capture as many positives as possible. For example: If we are building a system to predict if a person will not pay a loan, we want to capture it even if we are not very sure.

4. **F1 Score**
F1 Score (2*Precision*Recall/(Precision+Recall) combines precision and recall - used to measure a test's accuracy

Depending on the application, you may want to give higher priority to recall or precision. But there are many applications in which both recall and precision are important.
**It seems that F1 is the best metric to use for our task if we you want to predict a person who will not repay a loan, we also want to be sure that we are right ant this person will not fully pay (Precision) and on the other hand we also want to capture as many of people who will not pay (Recall) as possible. The F1 score manages this tradeoff."

Comparison table

	Algorithm	Model F1 score	Recall score	ROC_AUC score
0	Random Forest	0.905680	0.769063	0.855193
1	KNN	0.705382	0.008369	0.502907
2	Gradient Boost	0.906369	0.777376	0.858544
3	Naive Bayes	0.843023	0.880745	0.849342



Conclusion

I have evaluated each model in terms of model F1 score, Recall score and 'ROC-AUC' score for both the training and test data, and plotted them. The two best performing models are the **Random Forest** and the **Gradient boost**. Both are the ensemble model, based on decision trees.

Future Research

For future research I would like to create a model to predict if a person is qualified for a loan and determine loan classification grade.
I also would expand these models to include the data of applicants who were rejected by Lending Club.