# Loan default prediction

## Data from Lending Club


Group 17:

Elmira Nurbayeva

Xuân Tran

Shradha Khanal

Anastasia Hirvonen

Nguyet Nguyen

# Business Understanding

Lending club is an online peer-to-peer lending company located in San Francisco. It specialises in lending various types of loans to their customers.

There are two types of risks associated with the bank's decision when the bank receives a loan application:

- If the applicant is likely to repay the loan, then not approving the loan involves a risk of missing a profit
- If the applicant is not likely to repay the loan (default) then approving the loan would lead to financial loss

Therefore, the aim of our model is to help the company to accurately predict whether the customers are likely to default on their loans or not. It would help the company to improve profit and limit its losses.

The loan status has two classes 'Fully Paid' and 'Charged off'. We intend to use logistic regression with L1 penalty, and decision tree to predict which class an individual is likely to belong to.

# **Data understanding** - data set has a total of 27 variables

| Variable | Description | Variable | Description |
|---|---|---|---|
| loan_amnt | Loan amount | title | Purpose of the loan |
| term | The number of payments on the loan. Values are in months and can be either 36 or 60. | 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. |
| int_rate | Loan Interest rate | earliest_cr_line | The month the borrower's earliest reported credit line was opened |
| installment | The monthly payment owed by the borrower if the loan originates. | open_acc | The number of open credit lines in the borrower's credit file. |
| grade | LC assigned loan grade | pub_rec | Number of derogatory public records |
| sub_grade | LC assigned loan subgrade | revol_bal | Total credit revolving balance |
| emp_title | The job title supplied by the Borrower when applying for the loan.* | revol_util | Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. |
| 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. | total_acc | The total number of credit lines currently in the borrower's credit file |
| 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 | initial_list_status | The initial listing status of the loan. Possible values are – W, F |
| annual_inc | The self-reported annual income provided by the borrower during registration. | application_type | Indicates whether the loan is an individual application or a joint application with two co-borrowers |
| verification_status | Indicates if income was verified by LC, not verified, or if the income source was verified | mort_acc | Number of mortgage accounts. |
| issue_d | Issue date | pub_rec_bankruptcies | Number of public record bankruptcies |
| loan_status | target value: Fully Paid or Charged Off | address | Address |
| purpose | A category provided by the borrower for the loan request. | | |

# Data understanding - Variables

- Original dataset contained 396 030 observations and 27 variables (including the dependent variable)
- We investigated the data using visualisation to detect how each independent variable affect the dependent variable. In this report we demonstrate the variables that we believe should affect the prediction model the most.
- The charged-off (default) cases account for about 20% of all observations. Therefore, the distribution is imbalanced (Figure 1).
- Most of the loan amounts (for both fully paid and charged off) have a range between 5000 and 20000 (Figure 2).
- Customers are more likely to default on loans that have to be repaid over longer period of time, 60 months vs 36 months (Figure 3).

Figure 1

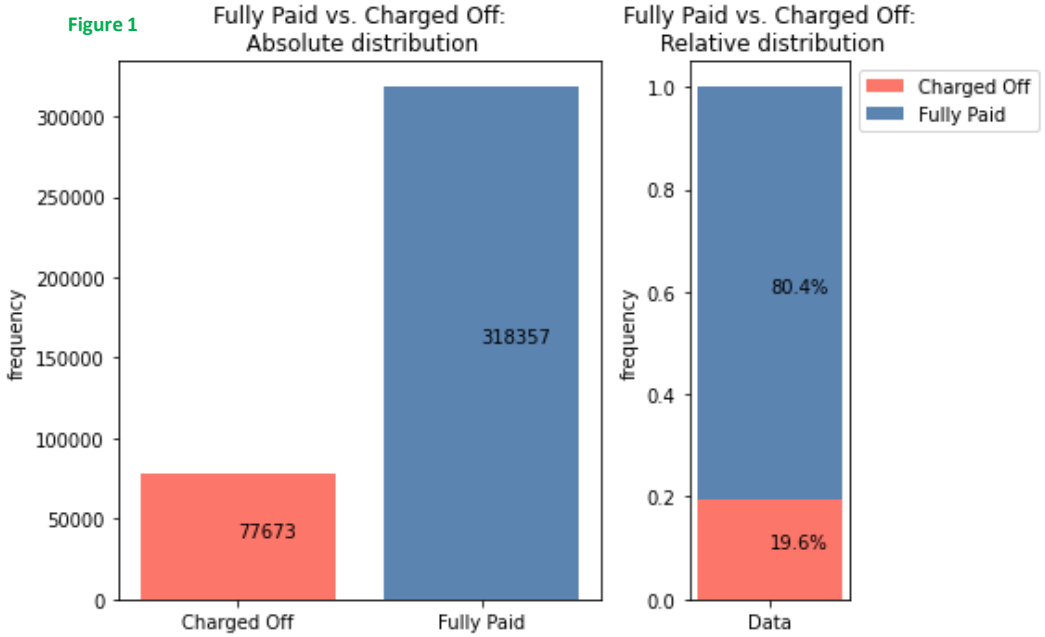



Figure 2

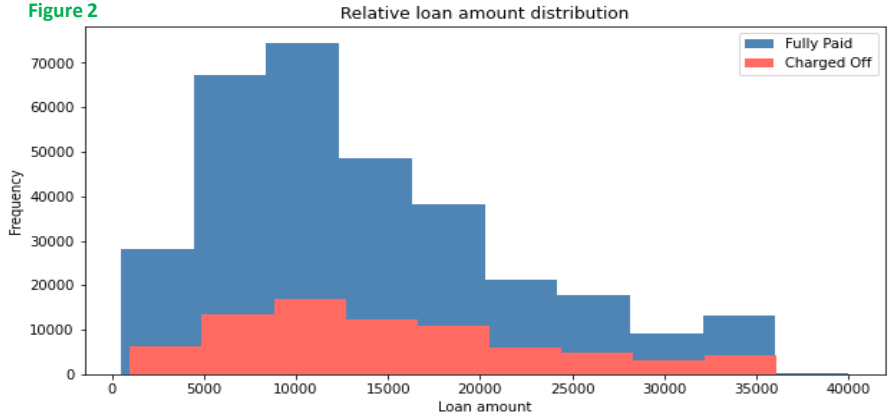
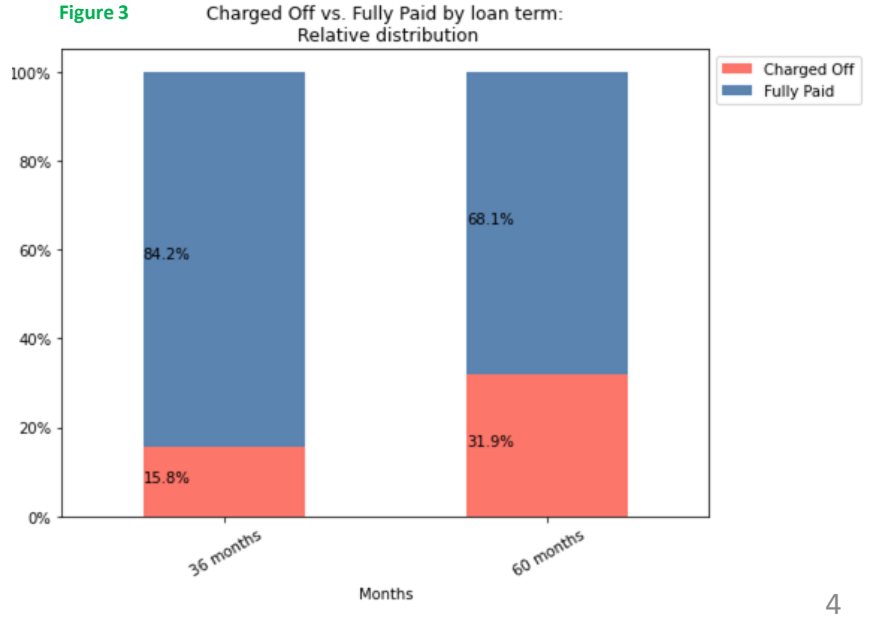



Figure 3



# Data understanding - Variables

Figure 4

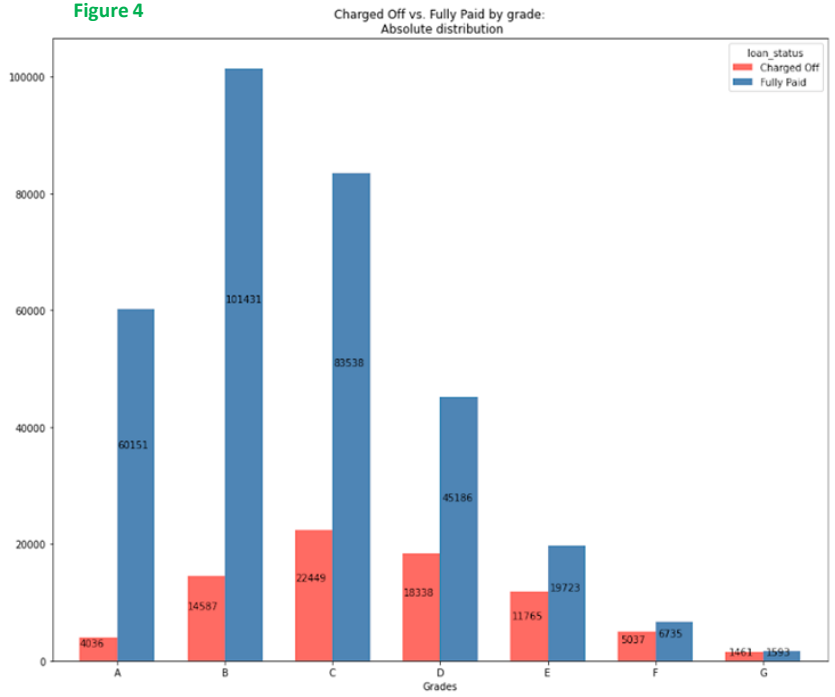



Figure 5

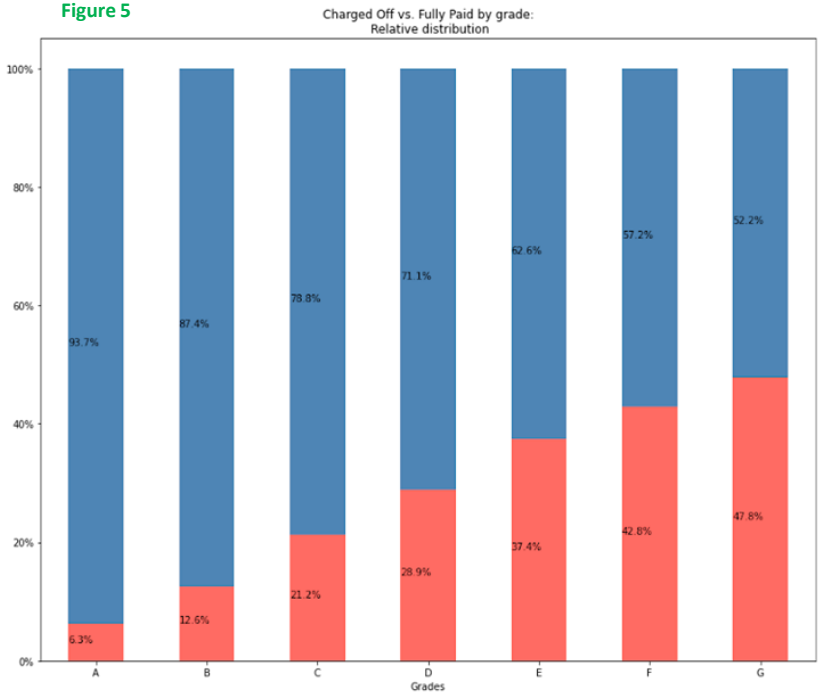
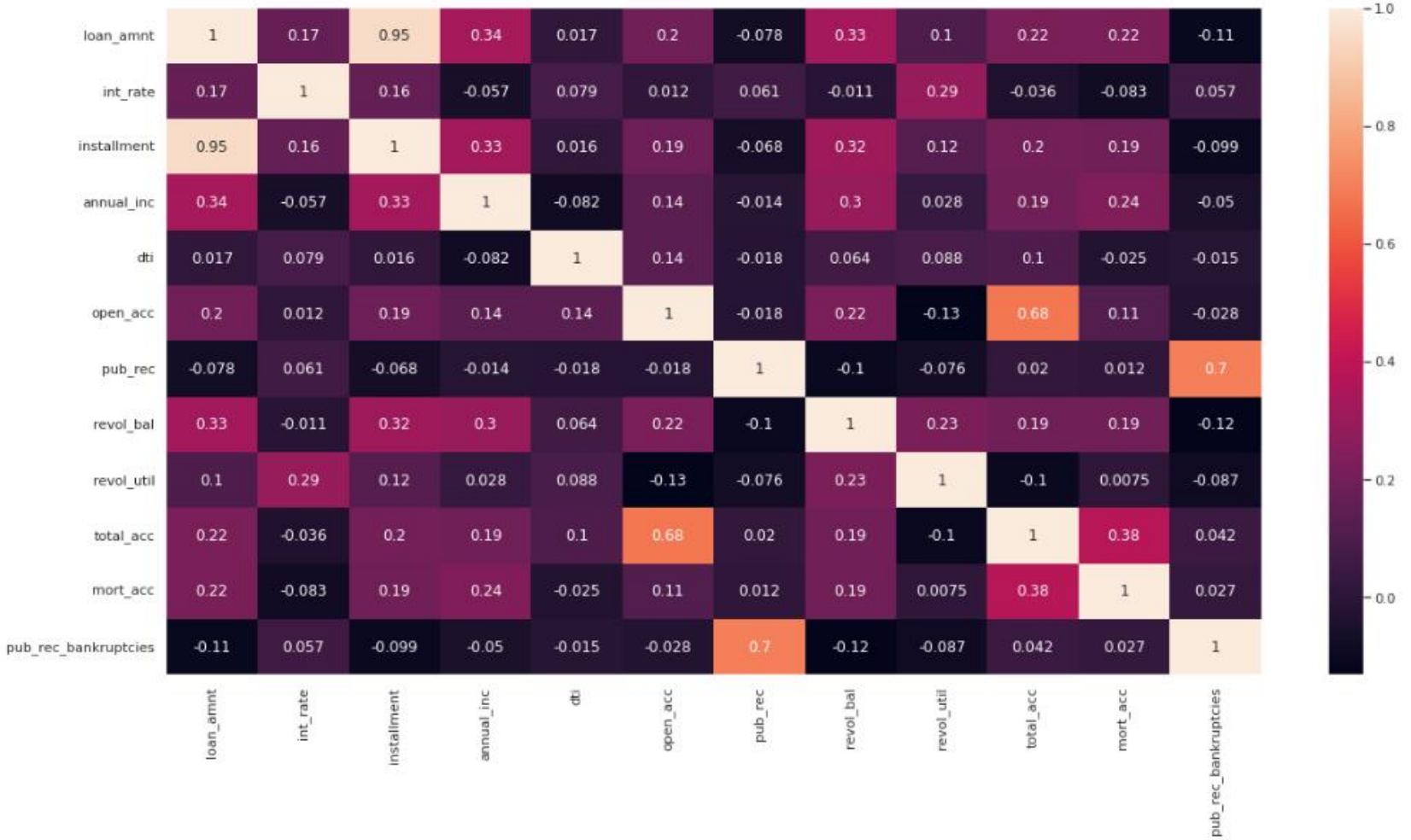



Figure 6


- 88% of all the observations are in loan grades A-D (Figure 4).
- The proportion of default on loans grows as the loan grade changes from A to G, with 47.8% of all loans not returned for Grade G (Figure 5).
- Zip codes 11650, 96700, 86630 have charged off (default) rates of 100% (Figure 6).
- We expect, that the status of the loan (dependent variable) will be highly affected by the zipcodes and loan grade.

# Data Understanding - Correlation matrix




**Some noticeable correlations:**

- Loan amount & installment: 0.95
- Pub-rec & pub-rec-bankrupcies: 0.7
- Total_acc & open_acc: 0.68

**Conclusion:** we choose only one variable from the correlated pairs of variables

# Data Quality and Preparation

- ✓ We conducted Exploratory Data Analysis to better understand our data [396 030 observations].
- ✓ Data doesn't have duplicate values, so no action needed
- ✓ We dropped some variables that don't contain useful data, e.g. title
- ✓ Correlation matrix showed that some variables are highly correlated, so we kept only one of correlated variables in order to avoid multicollinearity
- ✓ Data set had missing values, so after we decided on which variables we are using in our analysis, we dropped the rows with missing data [55 255 dropped, 340 775 observations left]
- ✓ We considered outliers values that are more than 3 standard deviations from the mean. We chose to drop those values instead of replacing it with some other value because there was relatively small amount of outliers and we have a large dataset. [14 764 dropped, 326 011 observations left]
- ✓ We applied robust scaler to scale the train/ test data because some numerical data are right skewed.
- ✓ Then we rebalanced the training data using SMOTE

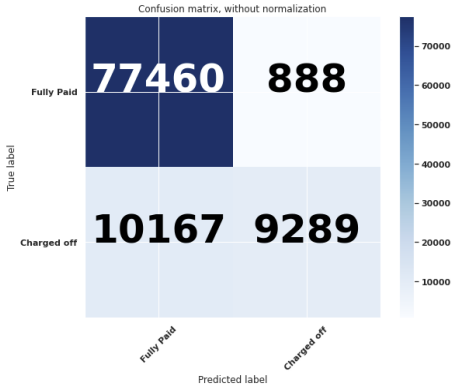
# Modelling

As a baseline model we run a Logistic Regression model with L1 (Lasso) Regularization Methods using an imbalance data. Further, we improve the same model by running it using the balanced data. As a comparison, we built a Decision Tree model using the balanced data as the third model.

Logistic Regression

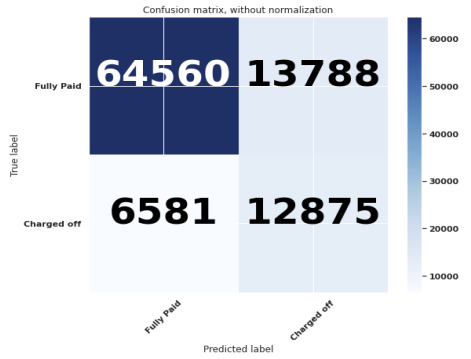
With imbalanced data

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.88 | 0.99 | 0.93 | 78348 |
| 1 | 0.91 | 0.48 | 0.63 | 19456 |
| accuracy | | | 0.89 | 97804 |
| macro avg | 0.90 | 0.73 | 0.78 | 97804 |
| weighted avg | 0.89 | 0.89 | 0.87 | 97804 |




With balanced data

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.91 | 0.82 | 0.86 | 78348 |
| 1 | 0.48 | 0.66 | 0.56 | 19456 |
| accuracy | | | 0.79 | 97804 |
| macro avg | 0.70 | 0.74 | 0.71 | 97804 |
| weighted avg | 0.82 | 0.79 | 0.80 | 97804 |

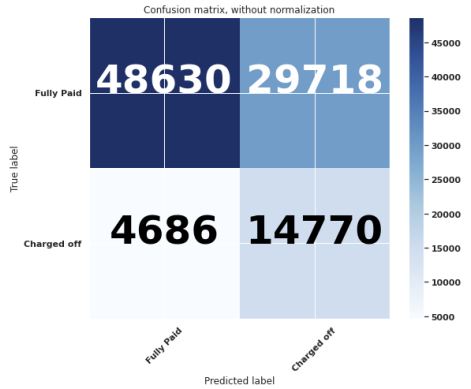



Understandably, the accuracy of the model decreases when we exploit the balanced data. But the recall has increased from 0.48 to 0.66,which means the model predicts better the defaulting loans.

Decision Tree

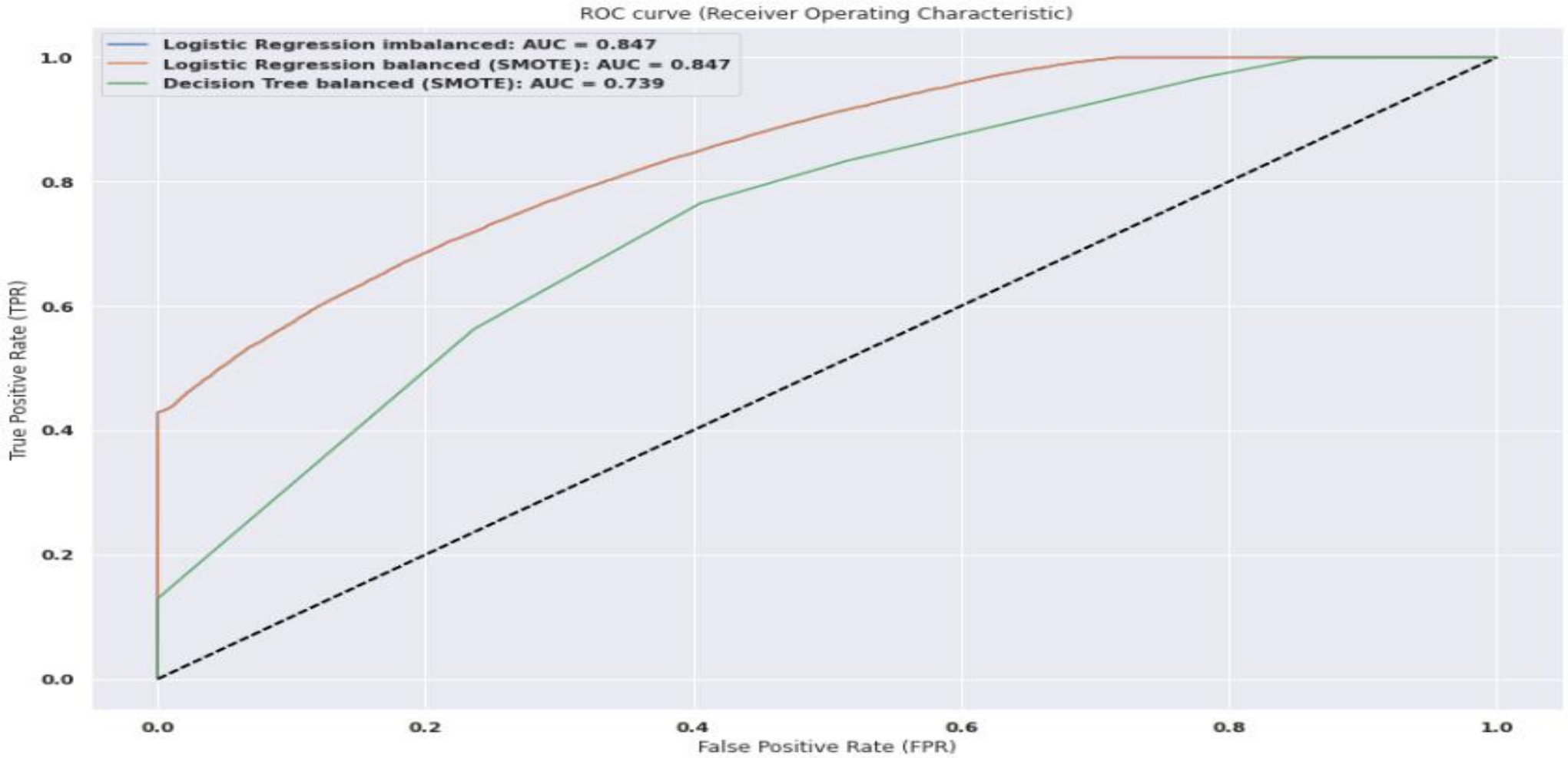
With balanced data

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.91 | 0.62 | 0.74 | 78348 |
| 1 | 0.33 | 0.76 | 0.46 | 19456 |
| accuracy | | | 0.65 | 97804 |
| macro avg | 0.62 | 0.69 | 0.60 | 97804 |
| weighted avg | 0.80 | 0.65 | 0.68 | 97804 |




Decision Tree model shows better results in detecting True Positive cases, but performs worse in True Negative. The models will be compared further using the AUC and expected benefit of the models.

# Evaluating models




We made ROC curves for all 3 models: 2 Logistic Regression and 1 Decision Tree.
Logistic Regression models have better AUC value (0.847 against 0.739).

# Expected Benefit

The expected benefit was calculated on an individual level.

Idea:

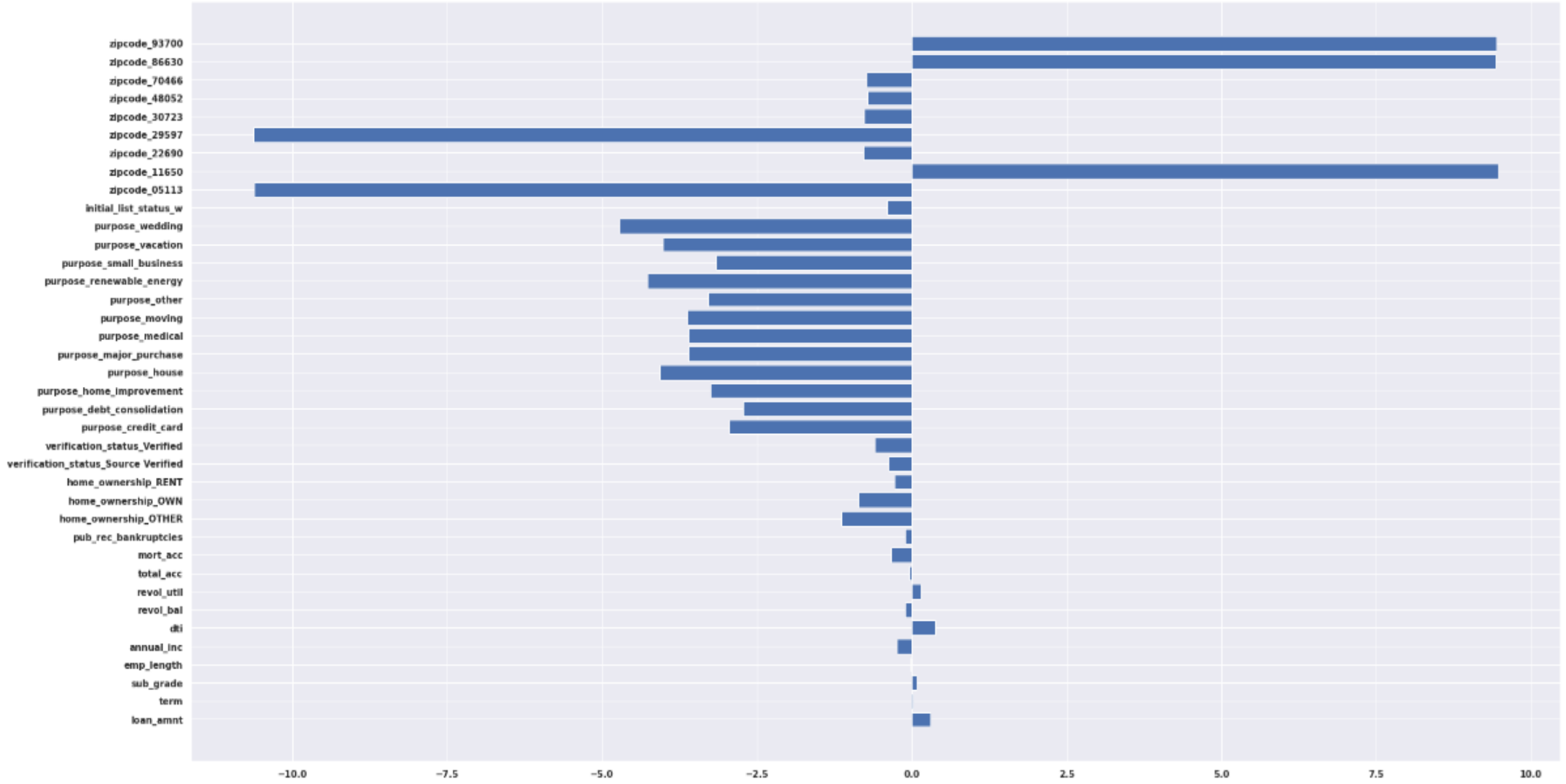
- If the model correctly predict that the loan will be returned (TN), then the benefit is the Gained interest = Loan amount X Interest rate
- If the model wrongly predict that the loan will be returned when in fact it is not (FN), then the loss is the Amount of the loan.

## Expected benefit:

| | Benefit (1) | Losses (2) | Expected Benefit (3)=(1)-(2) |
|---|---|---|---|
| Logistic regression with imbalanced data | 144 413 142.74 | -151 237 375.00 | -6 824 232.26 |
| Logistic regression with balanced data | 109 992 974.19 | -90 245 200.00 | 19 747 774.19 |
| Decision tree with balanced data | 81 962 631.14 | -65 159 300.00 | 16 803 331.14 |

According to the Expected benefit of the models, Logistic regression with balanced data preform better compared to the Decision tree model with balanced data

# Coefficients of Logistic Regression with Balanced Data




We extracted the results of the Logistic regression with balanced data. According to the model (that use Lasso Regularization Method) zipcodes of an applicant have a great effect in predicting the loan status.

# Model selection

Based on Expected Benefit, AUC area, and classification reports, the Logistic Regression with Balanced Data is chosen for the following reasons:

- On average, the company earns $ 19.7 million, the highest expected benefit among all models.
- AUC area has a high value of 0.847, the best AUC area among all models.
- The highest accuracy rate of 0.79 among models that use balanced data.

# Conclusion

**Some knowledge after building the models:**

- Balanced data supports better results.
- It is better to have different model types for comparison.
- Detecting outliers is important for model performance.
- Lasso regression improved the model, because only handful of features affect the target variable

# Recommendation

- We recommend to use the model Logistic Regression with Balanced Data in predicting whether the loan be returned or not. This model showed best performance among all others models based on having the highest accuracy, AUC, and expected benefit.

- Based on EDA, customers with zip codes 11650,86630, and 93700 have 100% probabilities of write-offs. So, we do not recommend to give a loan to an applicant with those zip codes.

- The features that affect the target variable the most are zip codes 93700, 86630, 29597,11650, 05113

- The grade of the loan has a significant effect on the target variable as well, as according to EDA, 94% of people who have grades “A” pay their loans on time. Therefore, the company should prefer the clients with higher loan grades.

# THANK YOU!