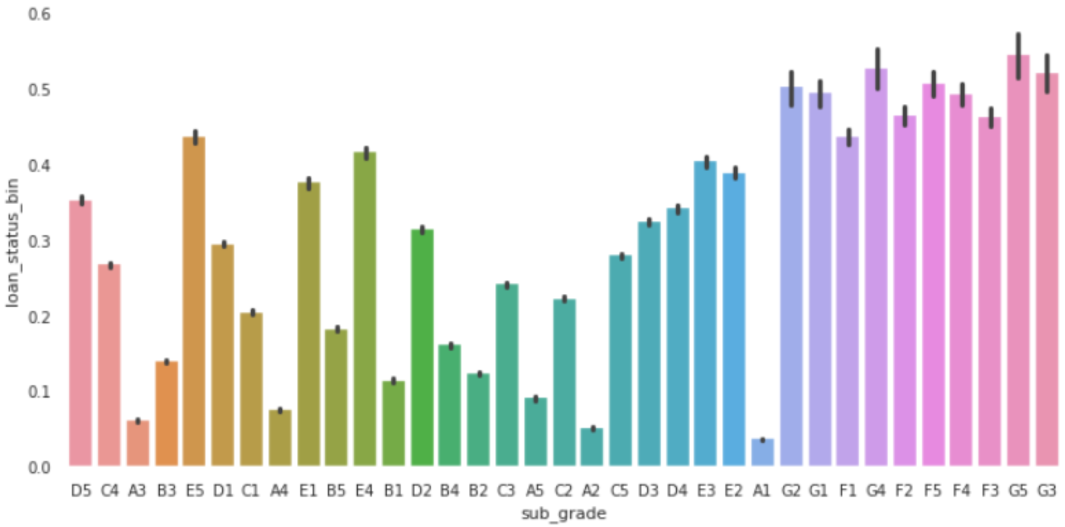
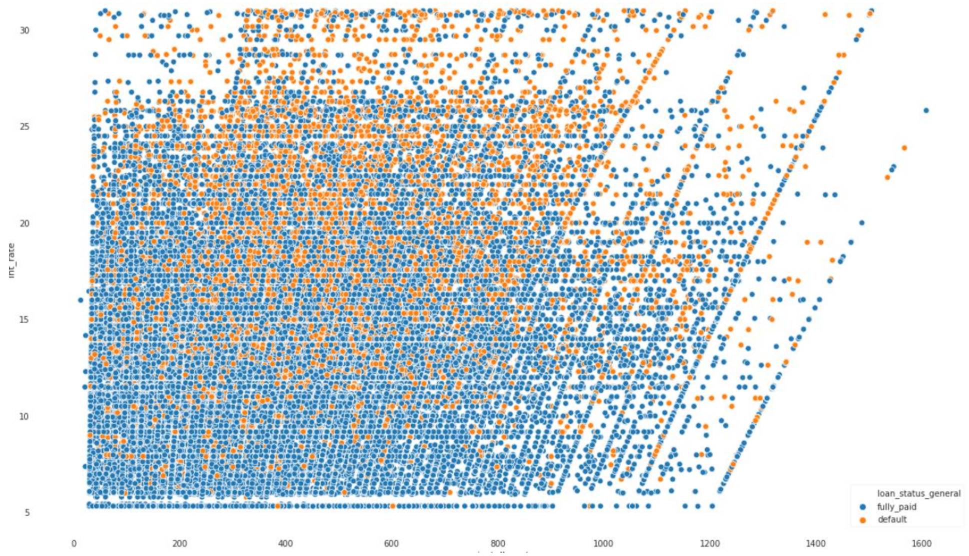
ML Task

Business and Data Understanding

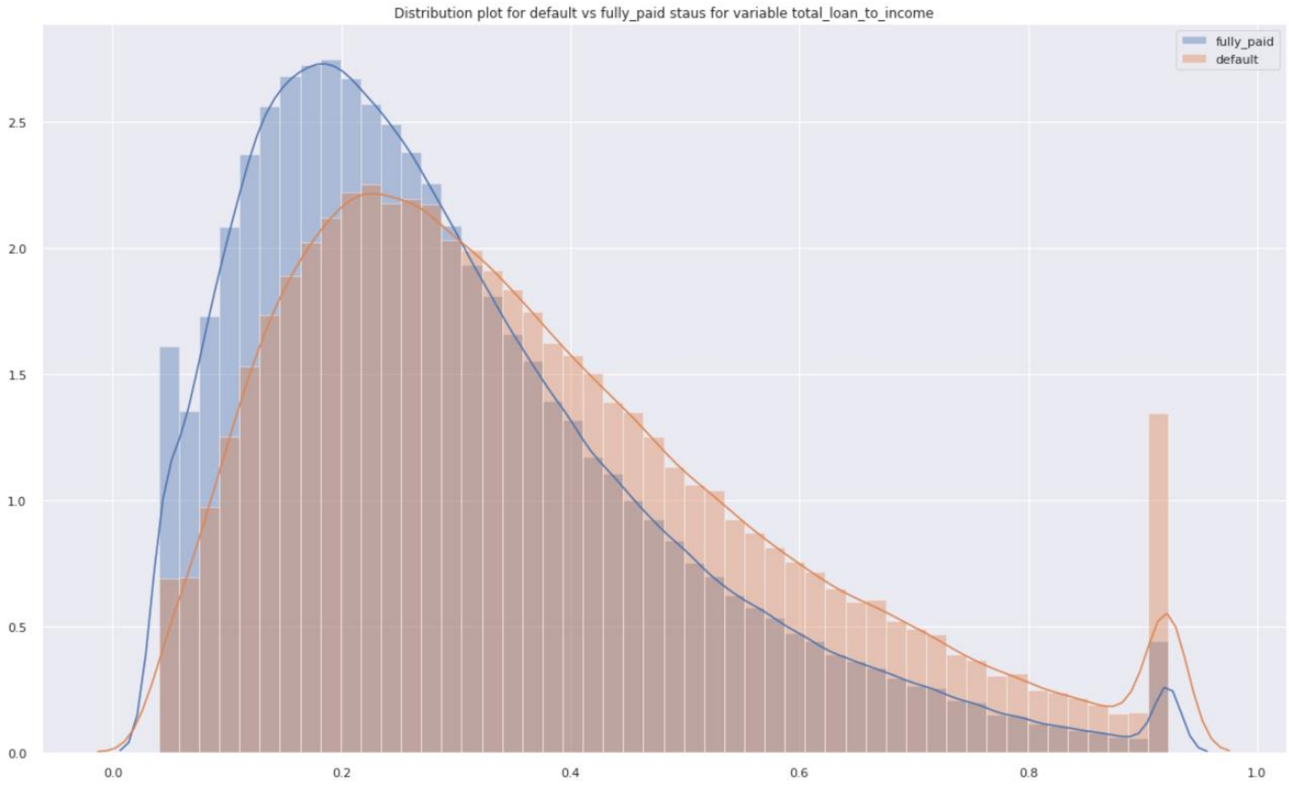
- 80% of work was used to preprocess data and understand business context.
- Check for null values, inconsistencies: missing imputation, dropping of leakage variables (ex: recoveries)
- Addressing the problem of predicting probability of default: filtering dataset to get only applications of interest.
- Dealing with self-reported variables outliers: using column capping concept to constrain outliers of these variable to avoid any kind of cheating.
- Multivariate distribution analaysis and correlation between variable to understand what is correlated with default applications.
  - Top: Rate of default within sub_grade applications.
  - Bottom: Instoptallment vs. int_rate. Interestingly, we can see see a cluster of default on right corner. This might be a confounder variable.

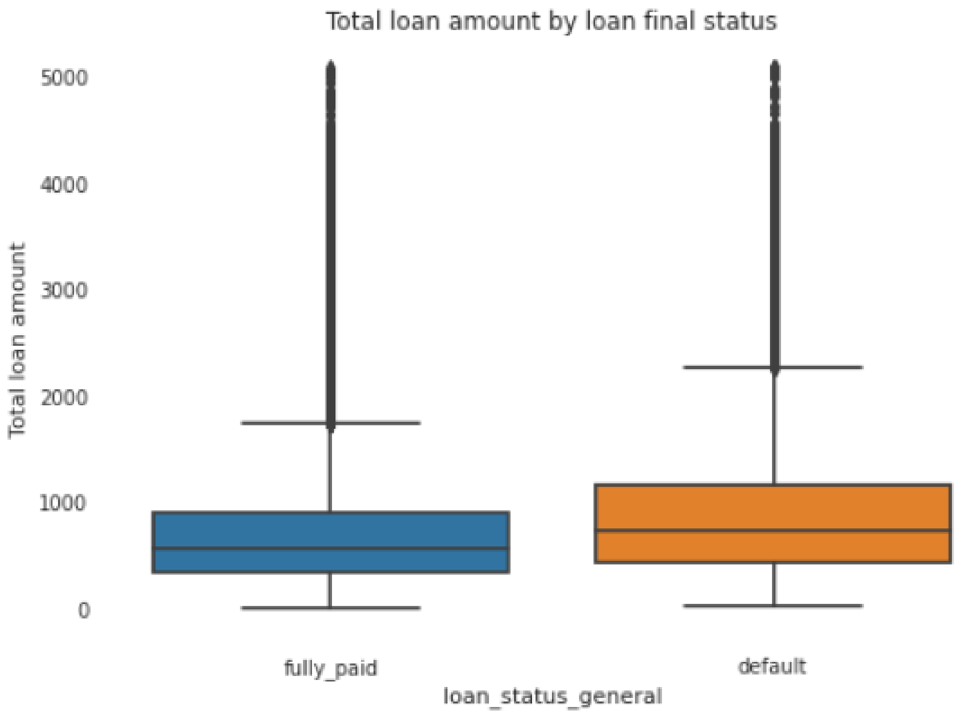




Data Preparation and Feature Engineering

- Construction of robust pipelines to apply preprocess steps devised on exploratory data analysis using scikit-learn API.
- Feature Engineering step to help models' performance: total_loan_amount, loan_to_annual_income rate (annual_inc capped to avoid outliers), etc.
- Modularized functions and classes in separate folder to facilitate and concentrate notebook code to analysis only.





Model Definition and Validation

DEFINING MODELS AND TECHNIQUES TO DEAL WITH IMBALANCE DATASET.

TESTING THREE MODELS: LOGISTIC REGRESSION, RANDOM FOREST AND GRADIENT BOOSTING. WHY? LINEAR, BAGGING AND BOOSTING MODELS TO DIVERSIFY PREDICTIONS.

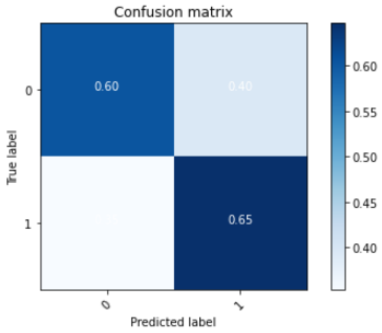
SAMPLING TECHNIQUES: RANDOM OVERSAMPLING, SMOTE, RANDOM UNDERSAMPLING, CLASS_IMBALANCE PARAMETER AND NEARMISS METHODS.

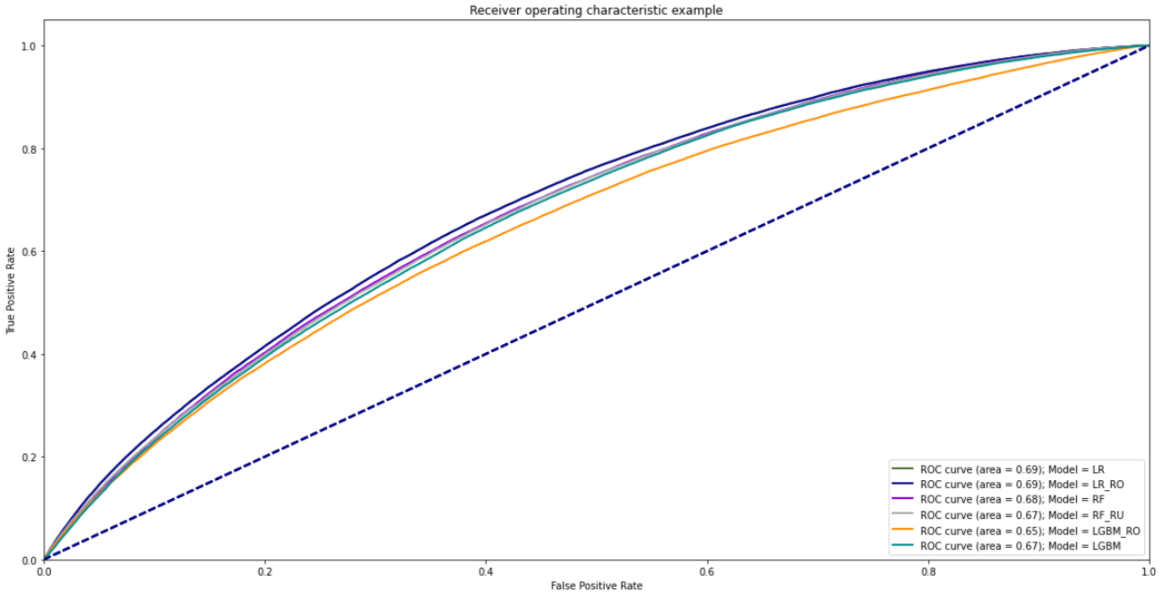
TRAIN-TEST SPLIT VALIDATION USING ISSUE_D. I DO THINK WE SHOULD NOT IGNORE TEMPORAL FACTOR IN THIS DATA, BECAUSE THIS IS ACTUALLY SIMULATING REAL WORLD.

HYPERPARAMETER TUNING USING TRAIN SET TO SELECT BEST MODEL’S PARAMETERS.

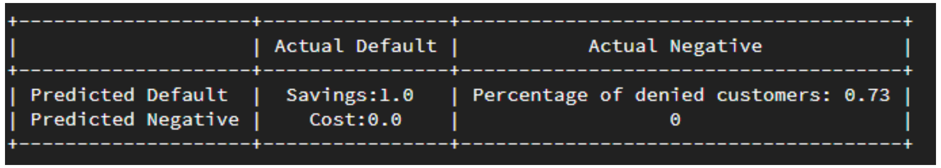
Result Analysis and Next Steps

- Analyzing precision-recall trade-off in this dataset. Usager of F1-Score to determine hyperparameters.
- ROC curves and discussion of threshold parameters in precision/recall metrics.





- Connecting machine learning metrics and business optimization. Cost matrix, how much Money are we saving / loosing with our predictions?



```
+-------------------+---------------+------------------------------------+
|                   | Actual Default |           Actual Negative          |
+-------------------+---------------+------------------------------------+
| Predicted Default  |  Savings:1.0   | Percentage of denied customers: 0.73 |
| Predicted Negative |   Cost:0.0     |                  0                  |
+-------------------+---------------+------------------------------------+
```

Next Steps:

- SHAP values, Partial Dependence Plots and Feature importante analysis to draw inference and interpretability of models outputs and where we could improve in feature engineering.
- More robust hyperparameter tuning (Bayesian or RandomSearch).
- Platt calibration for correct probabilities in tree models.
- Exploring models during payment of loan to assess if we could predict in which month will a borrower default.