CREDIT RISK ANALYTICS

PREPARED BY: MEGHANA RAO,DSP-08, MAY 2018

FLOW OF THE PRESENTATION

- Problem Statement
- The Approach
- Data Quality Issues
- Performance Report
- Business Opportunities
- Conclusion

PROBLEM STATEMENT

A credit scoring model is the result of a statistical model which, based on information about the borrower (e.g. age, number of previous loans, etc.), allows one to distinguish between "good" and "bad" loans and give an estimate of the probability of default.

The aim of the project is to build a credit scoring model that will help evaluate risk associated with a borrower.The model helps determine the health score of the borrower, determine the score of a new client and try to identify critical factors behind a borrower defaulting on loan.

A credit scoring model is just one of the factors used in evaluating a credit application. Assessment by a credit expert remains the decisive factor in the evaluation of a loan.

THE APPROACH

CRISP-DM METHODOLOGY

CHAMPION-CHALLENGER FRAMEWORK

- Implemented supervised machine learning algorithm.
- Since problem is about estimating the probability of default, Logistic Regression, Decision Tree and Random Forest algorithms were implemented.
- Grid Search Cross Validation to identify critical parameters and fine tune the models for optimum performance.
- Precision and Recall primarily used metrics to evaluate different models and choose the final model.

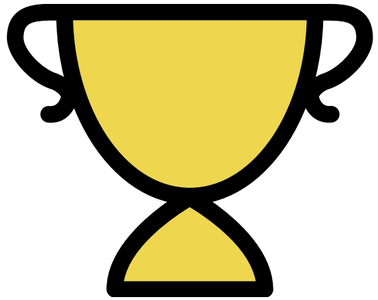
DATA QUALITY ISSUES

- Proportion of defaults- Imbalanced class
- Frequency of values in each variable
- Proportion of outliers
- Proportion of missing values
- Other issues like inconsistency,incompleteness

PERFORMANCE COMPARISION

High Precision and High Recall

High bias and low variance is preferred here



Logistic Regression

Precision-0.297
Recall-0.8572
Accuracy- 0.987

Decision Tree

Precision-0.582
Recall- 0.857
Accuracy- 0.995

Random Forest

Precision- 0.028
Recall- 0.91
Accuracy-0.819

BUSINESS OPPORTUNITIES

- A large number of people come from prominet areas of the US like California and New York where the cost of living is pretty high.
- A majority of people take loans to repay their previous loans so more offers can be clubbed together. One option could be lowering the interest rate in such a case or extending the deadline for payments.
- A very few people are on the higher side. So it can also be understood that some people may not want to disclose their income.

CONCLUSION

- Logistic Regression gave better results over the other two models. A choice for interpretability and visualisation due to its simplicity.
- There is no conclusive definition on what constitutes a default- the number of missed payments or the loan amount exceeding a particular threshold
- The missing values are imputed based on certain parameters there may be chance that the person has defaulted by accident and not due to any of the explanatory variables.

THANK YOU