FLIP ROBO

Micro-Credit Defaulter Model

Submitted by: Nipam Nayan Gogoi.

INTRODUCTION

Business Problem Framing:

Background:

A Microfinance Institution (MFI) is an organization that offers financial services to low income populations. Microfinance services (MFS) provided by MFIs could be Group Loans, Agricultural Loans, and Individual Business Loans etc. MFS become very useful when targeting poor families living in remote areas with little source of income, and no access to bank loans or even bank accounts in some cases.

Mobile financial services (MFS) are more convenient, efficient, and cost saving than the traditional high-touch model when delivering microfinance services. Even though the MFI industry are very useful in assisting low income families, the implementation of MFS have been uneven in both challenges and successes. Today, microfinance is widely accepted as a poverty-reduction tool, representing $70 billion in outstanding loans and a global outreach of 200 million clients.

The project:

This project aims to support one such client in the Telecom Industry. They are a fixed wireless telecommunications network provider. They have launched various products and have developed its business and organization based on the budget operator model, offering better products at lower prices to value conscious customers.

They understand the importance of communication and how it affects people's lives and their ability to support themselves financially. That's why they are collaborating with an MFI to provide micro-credit on mobile balances, to be paid back in 5 days.

The sample data is provided to us from our client database. The client wants predictions that could help them improve their future investments through the accurate selection of viable customers.

Review of Literature

Current global climate for cellular network providers:

Economies everywhere are currently contracting due to Covid19, including the economy of India. This will most likely lead to a reduction in the leverage by institutions and in the leverage that they are able and willing to extend to poor customers. For example, in the U.S., several large operators are deleveraging, including Verizon, AT&T, and Comcast. This is because there will be a greater need for free cash flow within institutions. (spglobal) Credit metrics will also be a risk. Economic contraction could negatively affect revenues, earnings, and cash flow from operators' business-to-business segment and lead to lower consumer spending on content, fewer connected devices, and migration to cheaper mobile and broadband plans. (spglobal) At the same time customers will have greater need of lending services due to reduced employment and bankruptcies of small and midsize enterprises. The risk of customers defaulting on repayments will also increase. (spglobal)

Technical research:

By utilising the three models, it is therefore possible to determine the most appropriate model structure for describing the relationships between the explanatory variables and default for airtime lending. First, logistic regression (LR) provides binary classifications using linear relationships (Cox 1958). LR is traditionally used as a relatively simple model structure and sets a benchmark for comparing the performance of the classifiers. Second, a decision tree (DT) is constructed to assess the potential improvement using a nonlinear model (William 1959). While offering the possibility of a nonlinear model structure, DT has the added benefit of resulting in a set of rules that are relatively easy to implement. Third, an ensemble approach known as Random Forest (RF) is deployed by averaging over a collection of decision trees (Breiman 2011). By including the RF model, it is possible to ascertain whether ensemble techniques can offer any benefits beyond the LR and DT techniques. (MDPI)

It is more important to predict those customers who will default than those who will repay. This is because the financial risk associated with defaulters is high. A confusion matrix is used to evaluate the classification models with positive (negative) outcomes denoting repayment (default), respectively. This 2x2 matrix measures the number of predicted/actual cases that are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). From this matrix, it is possible to calculate the classification accuracy. The accuracy formula is given by: $ACC = (TP + TN)/(TP + TN + FN + FP)$. The cost of one non-performing loan from a customer defaulting is much greater than the benefit of a customer repaying. The loss from rejecting a customer that will repay is much less than that suffered from approving a loan for a customer that will default. Accuracy is a useful summary statistic but is not the most relevant performance metric for this particular business application. The greatest threat to financial sustainability arises when the classifier predicts that a customer will repay a loan and they actually default (FP). Therefore, it is most important to correctly predict the customers who will not repay. For applications that require highly effective detection ability for only one class, it is recommended to consider an alternative metric to accuracy (Tang et al. 2009). The loan approval application is best assessed using the classification metric known as specificity defined as: Specificity = TN/(TN + FP) (2) Specificity measures the probability that a classifier correctly predicts default when considering all those that actually default. The priority for profitable lending is to avoid customers that are likely to default, which is achieved by maximizing specificity. (MDPI)

Interpretation:

Overall the client will want to be extremely sure that any loans it extends will be repaid. This is of course always the ideal but under current conditions it is crucial as the client needs to curb any additional losses.

This is complicated by the fact that their customers are facing their own challenges and the unstableness of the situation will cause a predictive model to over-estimate the likelihood of repayment.

Most providers use three variables to qualify users: ARPU (Average revenue per user); Tenure (i.e. how long the customer has been active on the network); Date of last call/top-up. (Solon)

Underwriting of the risk in microfinance has traditionally focused on the applicant’s capacity to pay, but is now it is now shifting towards a propensity-to-pay models that look up at the character of the applicant. (Wbank)

Model/s Development and Evaluation

1. Removed the categorical columns:- msisdn - mobile number of user, pdate - date, pcircle - telecom circle since these are irrelevant for the Analysis.
2. Outliers are removed based on Z-score.
3. The Dataset was Unbalanced. The label with 1 was 7 times that of 0. It was treated using Oversampling.
4. Further the data was scaled using Sklearn Standard Scaler before Model Building.
5. Used 4 models :- Decision Tree, Gaussian Naïve Bayes, Logistic Regression and Random Forest from Sklearn module out of which the best performance was by Random Forest.

Conclusion

1. The data contained outliers which was dealt using Outlier removal before Prediction.
2. It was an Unbalanced Dataset which was dealt using oversampling.
3. Based on the heatmap many columns are somewhat correlated to the target column, but none are strongly correlated.
4. Decision Trees and Random Forest, which is simply a collection of decision trees performed best on the dataset. The precision is accurate with zero False Positive(FP) values but recall is little bit less with a few number of False Negative(FN) values.
5. The Best model performed 98% accuracy on the unseen Test Dataset.