❖ Project Title: Designing 'Zero credit touch' (ZCT) pre-approved credit underwriting program for retail customers

❖ Name of the Project Guide: Sandipan Ray

❖ Name of the Project Buddy: Ankur Khandelwal

Brief description of the project:

The bank has developed various ZCT strategies where without any credit intervention and additional information taken from customers, credit facilities can be provided. Objective of this internship is to further improve these models and develop data analytics based solution for each of the problem areas identified.

1. Development of credit underwriting model considering all possible information available about the customers within the bank
2. Estimating income for customers where it is not available so that offer amount appropriate for a customer can be given.
3. Advanced machine learning / deep learning approach should be explored for developing potential fraud identification model and should be able to flag off majority of the known fraud cases

❖ Relevance of the project:

The Bank is focusing on sourcing a significant amount of retail credit card and asset business (personal loan, home loan and auto loan) sourcing primarily from internal customers to reduce credit and operations cost of lending whilst providing a superior customer experience. Any improvement in these models will have a significant impact on both top line and bottom line of the Bank.

❖ Scope of the project:

The bank has developed various ZCT strategies where without any credit intervention and additional information taken from customers, credit facilities can be provided. We have identified following challenges in the expansion of ZCT strategies

1. Credit models which are combination of business rules, scorecards and machine learning models, do not qualify a significant proportion of existing ICICI Bank customers.

2. Wherever customers do not have a salary account with the Bank, estimated income is lower leading to customer being offered an amount lower that his/her requirement
3. Customers with fraudulent intension has opened accounts and overtime has created profile which would qualify for ZCT. Although number of such incidence are less, still it can become a material cause for credit loss once the ZCT strategies expands

Objective of this internship is to develop data analytics based solution for each of the problem areas identified.

1. Development of credit underwriting model considering all possible information available about the customers within the bank namely savings bank account transaction data, credit card behavior, repayment track information from credit bureau and other profile related information captured in various CRM / internal platforms. Model(s) developed should performance consistently across various product portfolios and across various time horizons.
2. Estimating income for customers where it is not available so that offer amount appropriate for a customer can be given. However estimated income should not significantly overestimate the income which needs to be demonstrated on a holdout sample
3. Since there are only limited number of identified fraud cases, currently potential fraudulent customer identification is basis business rules. Advanced machine learning / deep learning approach should be explored for developing potential fraud identification model and should be able to flag off majority of the known fraud cases

Methodology

1. The bank currently uses 4000+ derived features to train credit underwriting models which are a combination traditional scorecards, machine learning model (XG Boost) and business rules. This internship should improve upon the existing models by incorporating following suggested methods.
   a. Along with existing features, automated feature creation techniques like Locally linear embedding (LLE), Topological data analysis (TDA) or similar methods in representation learning literature should be explored to increase predictability. Rather than developing one single model, exploratory data analysis should be carried out to identify segments which require separate models.
   b. Use of deep learning techniques especially recurrent neural network (RNN) class of models like LSTM for customers with long history of data available.
2. Estimating income for customers where it is not available. Currently models are built based on samples where verified income is available. Various techniques like regression machine learning (XG Boost) has been attempted however the performance of the models are still not at a desirable level. In this internship, deep neural network like multi-layer feed forward network (MLFFN) models should be explored along with a segmented approach basis exploratory data analysis. Segment of customers where income estimation is not possible / are prone to larger variance should be identified.
3. Since there are only limited number of fraud cases, Generative adversarial network (GAN) class of models can be used for over sampling identified fraud cases. Auto

encoder and it's variant should be explored for developing potential fraud identification model.

❖ Key deliverables of the project:

1. Credit underwriting model(s)
2. Income estimation model(s)
3. Fraud detection model(s)

❖ Project Location: Toronto

Intern will be based in Toronto, working remotely with ICICI Bank India. He/she will be assigned one data scientist for accessing all required data for analysis, helping with any logistics issues, explaining the analytical work environment etc.

❖ Relevant pre-read material that the intern should refer to in order to be well-prepared for the internship:

The intern should be hands on with various machine learning platforms like R and python. It is desirable that he/she is comfortable working in Spark and Google cloud environment. In terms of analytical techniques, theoretical and implementable knowledge of following techniques is a must
1. Machine learning – XG Boost, Random forest, isolation forest algorithm
2. Deep learning – Multi layer feed forward network, recurrent neural network, auto encoder and its variants, Generative adversarial network
3. Other topics – Locally linear Embedding, Topological data analysis