## "PROJECT PROPOSAL – PRUDENTIAL LIFE INSUARANCE ASSESSMENT"

<u>Problem Statement</u> – In today's world, people are more conscious towards health and tend to opt for various life/health insurance policies. One such provider in Prudential Life Insurance Policy.

In order to buy insurance, customers provide extensive information to know the eligibility and to identify risk classification to the providers. They also undergo medical exams which can take an average of 30 days. Because of this time-consuming approach, customers tend to lose interest.

<u>Scope</u>- To develop a predictive model to classify high risk and low risk applicants in order to make the process of identifying target customer segment quicker and less labor intensive while maintaining privacy boundaries.

## **Steps involved in process:**

Data Collection

- <u>Data Sets</u> We have collected a data set with 128 categorical, continous and discrete variables relating to applicant's personal information, medical history, family's medical history, employment history and insurance needs.
- <u>Inclusion of dummy variables</u>- A set of dummy variables relating to the presence of/absence of a medical keyword being associated with the application is also included in the data sets.
- <u>Source</u> We have collected the training set with 59,381 instances and test set with 19,765 instances with 128 attributes from Kaggle. (https://www.kaggle.com/c/prudential-life-insurance-assessment/overview)

Data Preparation

- Data leakage Excluding variables with less immportance
- •Correlation Omitting various attributes which have correlation > 0.8
- Missing values Replacing all NAs in the data set with proper values like medians.
- •Splitting training data divided into training and validation sets i.e. 85:15. Performance will be measured on test data.

Data Analysis and Reporting

- •Tools R programming
- Models Decision tree/ Random Forest
- •Generalised Linear Models (GLM), Gradient Boosting Machine (GBM), Xgboost
- Support Vector Machines (SVMs)
- •Finally selecting best model to predict applicant's risk on the basis of confusion matrix, ROC curves, AUC calutions and Lift charts.
- Error analysis

<u>Target variable</u> – "Response" is an ordinal measure of risk that has 8 levels. We will be categorizing our applicants in "High risk applicants" – should be given priority for life insurances with responses >4 and "Low risk applicants" – less priority with responses <= 4. Hence, it will be a binary classification problem.

<u>Milestone</u>- Till the next submission, we are planning to focus on data preparation and exploration with building models with decision tree, random forest and GBM methods.

**Team Members:** Ashok Bhatraju – 670248723, Nikita Bawane – 661069000, Ritu Gangwal – 670646774