Statistical Data mining Project

Prudential Life Insurance Risk Assessment

Karun Parashar
Jeetendra Gan
Chandana Singh

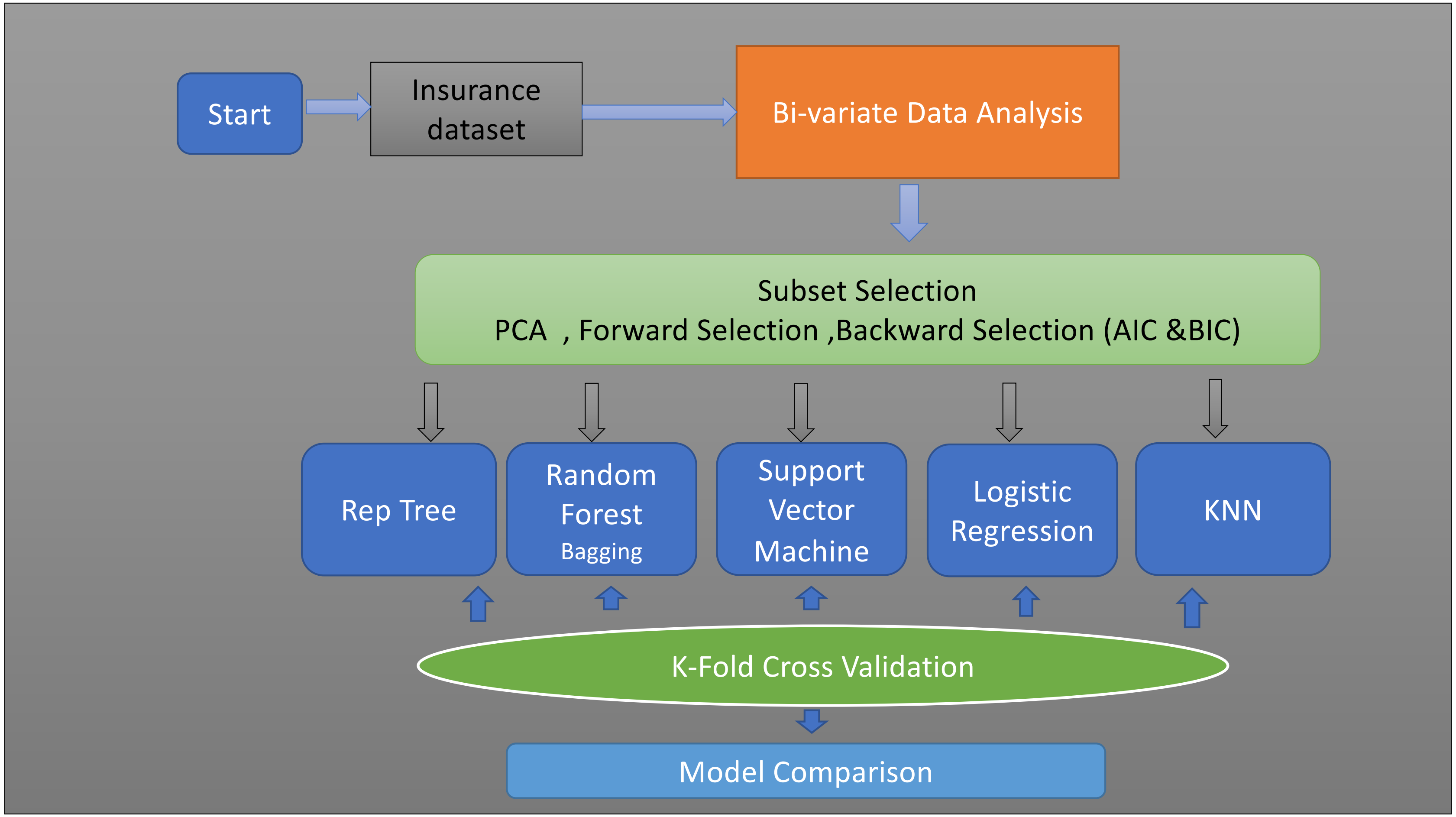
Introduction

- This project targets on building a Life insurance predictive analytics on underwriting process which involves potential customer risk assessment analysis.
- Involves risk classification of a new customer based on their personal details as well as medical and financial history.
- Improve decision-making and make the underwriting process faster and more economical.
- Helps the life insurance business to enhance customer acquisition and customer retention.

Risk Factors in Life Insurance

- The various risk factors to be assessed prior to underwriting decision are:
  1. Medical
  2. Financial
  3. Personal

Project workflow



Data Description

- The data set consists of 59,381 applications with 128 attributes, which describe the characteristics of life insurance applicants
- The data set comprises of nominal, continuous, as well as discrete variables, which are pre-processed and anonymized
- The values for each column are in a range 1-10, all being integers
- The response variable indicates the risk associated with each predicted value and is an 8 class response as classes 1-8 with risk decreasing from 1 to 8th class

Data pre-processing and Exploratory Data Analysis

- Handling of NA values as follows
- The threshold of NA count is set to 10% of data which approximates at 6000 and any column above this will be marked off
- Around 12 columns satisfied this criteria and were deleted
- Another column contained NA values but had a count of mere 19, hence all subsequent records associated to those values were removed
- Bivariate Analysis was performed using correlation matrix to find highly correlated predictors and calculate redundancy

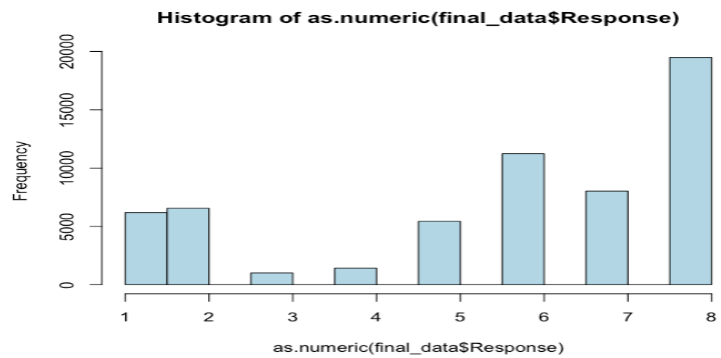
Subset and Feature selection

- Since the number of predictors is very high, the number is reduced using stagewise subset selection
- After performing forward and backward stepwise selection, the AIC,BIC and R square values were compared and the best variable model was found to be a 35 variable model which was inline with the bivariate analysis previously performed
- The model fitting was performed on the reduced model

Class Balancing

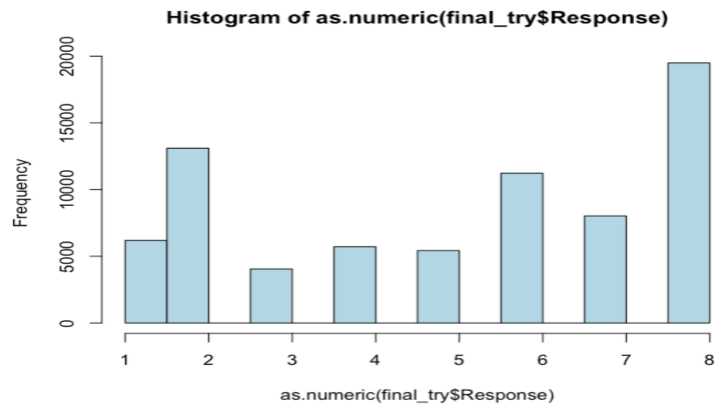
- The following histogram shows that the classes 2 and 3 are devoid of much data points

Pre Balancing



- This data was balanced using oversampling technique and now the distribution is as follows

Post Balancing



Non Ensemble Model Performance Statistics Unbalanced Data

- Logistic Regression-Polynomial and Multinomial regression

Model Selected	MSE(Mean Squared Error)	Classification Accuracy
Polynomial Regression	6.452	42.84
Multinomial Regression	5.998	47.79

- Support Vector Machines- Linear and Radial Kernel

Kernel Selected	MSE(Mean Squared Error)	Classification Accuracy
Linear	5.729	47.60
Radial	5.601	49.47

Non Ensemble Model Performance Statistics Balanced Data

Support Vector Machines- Linear and Radial Kernel

Kernel Selected	MSE(Mean Squared Error)	Classification Accuracy
Linear	6.109	44.71
Radial	5.872	48.04

- Support Vector Machines- Linear and Radial Kernel

Model Selected	MSE(Mean Squared Error)	Classification Accuracy
Polynomial Regression	7.985	41.14
Multinomial Regression	6.809	45.33

- This indicates that class balancing results in worsening of the classification accuracy for non ensemble methods

Ensemble Model Performance Statistics Unbalanced Data

- Rep tree(Growing and pruning a tree)

K-fold Validation	MSE (Mean squared error)	Classification accuracy（%)	Time Taken build a model(seconds)
5 fold	6.329	49.94	28
10 fold	6.331	50.29	30
20 fold	6.328	50.27	60
30 fold	6.330	50.30	140

- Random Forest

Number of Trees	MSE(Mean Squared Error)	Classification Accuracy
50	5.955	52.31
1000	5.988	53.06
2000	5.987	53.10
5000	5.983	53.24
10000	5.982	53.25

- Bagging(Number of Trees is set at 1000)

M Value selected	MSE (Mean squared error)	Classification Accuracy
6	5.954	53.06
7	5.889	53.06
8	5.960	53.08
9	5.977	52.73
10	5.989	52.75

Ensemble Model Performance Statistics Balanced Data

- Bagging (number of trees is set at 1000)

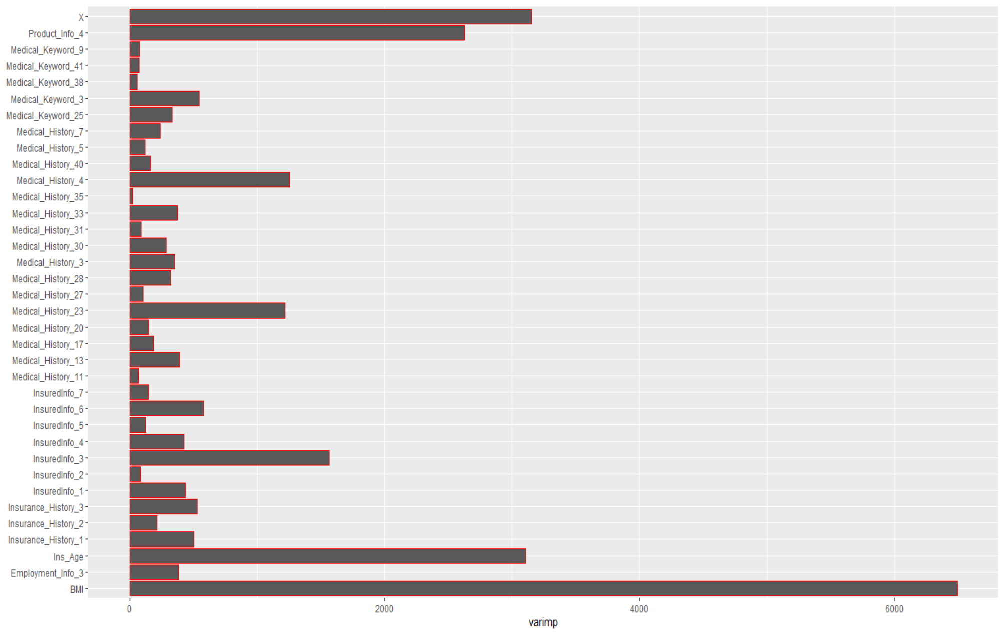
M Value selected	MSE (Mean squared error)	Classification Accuracy
6	4.882	58.74
7	4.821	61.11
8	4.331	64.05
9	4.112	65.07
10	3.987	65.72

- Random Forest

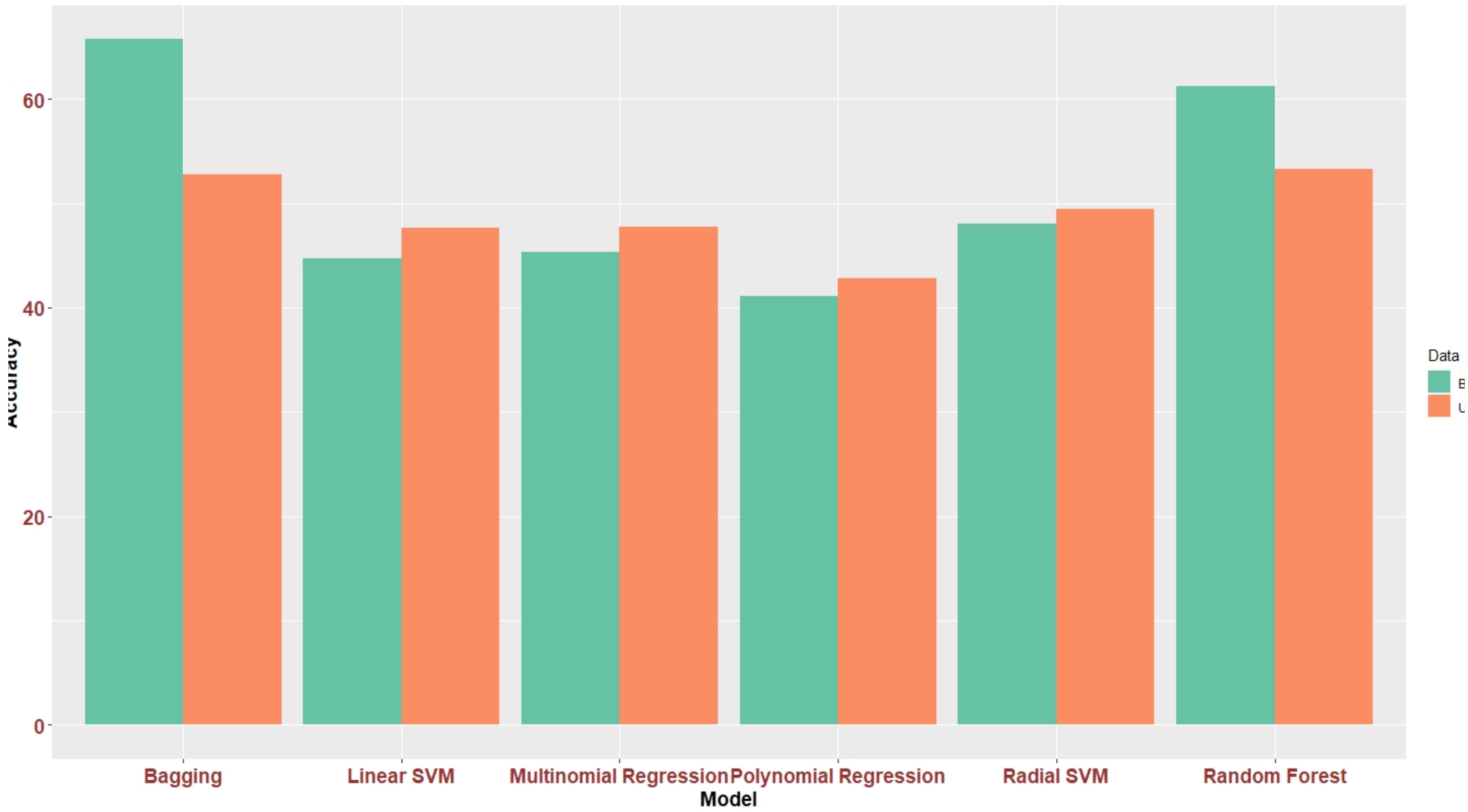
Number of Trees	MSE (Mean squared error)	Classification Accuracy
50	4.999	57.44
1000	4.912	58.82
2000	4.839	59.00
5000	4.832	60.11
10000	4.812	61.22

- The reason we are only considering the Bagging and Random Forest methods for the balanced data is because we got the maximum classification accuracy using these methods for unbalanced data
- As it can be seen the performance has substantially improved after class balancing which implies ensemble methods are the way to go while dealing with class imbalance using sampling techniques

Variable Importance graph (Random forest)



BMI is the most important variable followed by Age .



Accuracy
60
40
20
0
Bagging
Linear SVM
Multinomial Regression
Polynomial Regression
Radial SVM
Random Forest
Model
Data

Conclusion

- Ensemble methods improves the overall performance in terms of classification accuracy and mean squared error as compared to non ensemble methods.
- Class balancing enhances the accuracy by 10 % which is quite significant when compared with pre and post balanced train data model performance on Random forest and bagging.
- Model interpretation is excellent as the trees and the variable importance graph shows BMI and Age as the most important predictors.
- Dimension reduction from 116 to 35 predictors decreases the model complexity without any significant performance degradation
- Ensemble model with reduced dimensions using class balancing gives us the most parsimonious model