**Executive Summary**

Banks earn a major share of their revenue from the interest on loans, but this is often associated with a high level of risk. One of the most significant challenges that banks face is loan defaults. Banks are using forecasting techniques to analyze patterns that are more likely to be present in loan defaulters in order to mitigate this problem. Our dataset is made up of borrowers' past data collected by banks. Our aim is to identify a classification model that could be used to classify borrowers as likely to default or not.

The dataset we chose for our analysis comprises of 148,670 rows with 34 attributes. The attributes describe the nature of the loan and of the borrower. As our first step in the analysis, we excluded missing values, searched for outliers, and even excluded columns that did not contribute to our class of interest. The multivariate analysis using scatterplot matrix also showed some interesting correlations between the predictor variables. Borrowers with a higher property value, for example, are able to claim a larger loan amount. In addition, a higher loan to value ratio (LTV) increases the likelihood of loan defaults. Similarly, a higher debt-to-income (DTIR) ratio suggests a greater risk of default on a loan.

We then tried the classification models on our dataset to determine which model performed the best for our analysis. We compared important metrics for model selection, such as AUC, Accuracy of 1s prediction, Lift, and so on. There are 21 categorical variables and 7 continuous variables in our input variables. Ensemble Model 1 performed well under ROC, with area under the curve being 0.8094, a misclassification rate of less than 13%, and the best accuracy of 1s prediction (83.55%). The Ensemble model was developed by averaging logistic regression, boosted trees, neural networks, boosted forests, & the Naive Bayes Model, resulting in a robust model in nature. Importantly, since each model used is unique, Ensemble model reduces the variance in the predictions, resulting in improved predictive performance.

Finally, we concluded from our data analysis that the loan defaults occurring for business accounts were more as compared to non-commercial loans. Higher LTV values corresponded to high default chances and thus a threshold can be set in order to process future loans. The banks could utilize this analysis to restructure their loan lending process which can be further expanded to other financial systems.

Default in financial systems indirectly impacts the country’s economy as banks function on a large scale. Hence, it is essential to identify loan defaults to optimize the bank’s profits which could eventually benefit the economy as a whole.