DS Case Study

Acme Bank is an internationally renowned bank which has been in the Indian banking sector since the past two decades. In the last year, they have started offering a premium travel credit card in association with AirIndia. Typically, the credit card business has been a profitable one for Acme. However, the Acme travel card has not registered profit so far. This is a worrying trend for the CEO as the overall Acme portfolio is stagnant and they were hoping to find new customers by targeting a more premium clientele.

A closer evaluation of the balance sheet shows that an average of 20% revenue is being lost monthly due to credit card default. If this revenue leakage is stopped, then the CEO believes that this card could become one of the most profitable product offerings in the company's portfolio.

As part of a specialized analytics firm, you are called in to assess if the cardholders' past data and history can be used to predict default. If one is able to predict default even a month before its' occurrence, significant revenue savings can be made

You are provided with a consolidated database of individual customers and their credit history (refer to the file sent by email, *"acme_bank_data.csv"*)

The data team has also provided you with a data dictionary (*"acme_bank_data_dictionnary.csv"*)

To get started, do the following:

I. Import the dataset into R
II. Take a look at the dataset, what would be some of the first checks you would do?

III. In the dataset, are there any missing values? How will you treat them?
Hint : Think about imputation versus dropping the values

IV. What about outliers, which methods can be used to check for outliers. What should be done with these outliers?

Remember to run your code and get a dataset with missing values and outliers treated

V. Looking at the dataset, can you suggest any new features/variables which can be created to help in predicting default? Make at least 3 new features to be used for the model.

VI. Now after completing basic data treatment and feature engineering, you start thinking about prediction of default. Before diving into algorithm building, are there any visual/graphical methods you can use to analyze the relationship of default with other variables in the dataset.

VII. Based on this, are there any key variables which emerge for prediction of default?

VIII. Which algorithms can you use to predict default?

IX. Let's set up a basic logistic regression model. Should we use all the variables while making the model? Why or why not?

X. Which variables will you consider for building the model?

Use the shortlisted variables to make a logistic regression model. Based on your results, answer the questions which follow.

XI. Is the model significant

XII. What is the most predictive variable?

XIII. Is gender significant in predicting default? Is this different from your preliminary findings? What could be the reason for this?

XIV. How will you measure the performance of the model?

XV. Based on the confusion matrix, what is the true positive rate of the model

XVI. The logistic regression model performs well in capturing default, but its' sensitivity to non-defaulters can be improved further. How can we make this improvement?

XVII. Run a CART tree on the same data. What is the AUC of the model now? Think of why this change has occurred.