**Glossary**

|  |
| --- |
| 1. **Dataset Predictors and Labelled Class**   *City, StateProvinceName, CountryRegionName:* Residential area of each customer  *HomeOwnerFlag:* Whether customer owned his/ her house  *AveMonthSpend: Average monthly amount customer spend in the company*  *BikeBuyer: Labelled class showing whether the customer has purchased a bike in the company* |
| 1. **Evaluation Methodology & Performance Measurements**   *Cross-Validation:* A resampling procedure used to evaluate machine learning models on a limited data sample by splitting data into k folds, train on k-1 fold and validate in remaining fold for k times. *F1 Score:* Covey the balance between Recall and Precision by formula {2\*((Precision\*Recall)/ (Precision+ Recall))}, where recall is the true positive rate and precision is the predictive power of model.  *PRC (Precision-Recall Curve):* summarize the trade-off between the true positive rate and the positive predictive value for a predictive model using different probability thresholds  *Area Under Curve (AUC) of PRC:* Area Under the PRC curve, which suggests a summary of model skills. |
| 1. **Naïve Bayes**   *Model Definition:* A simple classifier that built upon Bayes’ rule of conditional probability (Assume class prior probability) and calculate the posterior probability. It adapts a naïve approach that all predictors are conditionally independent  *Bayes’ rule of conditional probability:* To calculate the probability of an event (named as posterior probability) given another prior event(s)  *Prior Probability/ Class prior:* Probability of the prior event. In this study, it refers to probability of the labelled class  *Zero Conditional Probability:* If an instance in test data set has a category that was not present during training then it will assign it “Zero” probability and won’t be able to make prediction.  *Additive smoothing:* A technique used to smooth categorical data by adding a pseudocount to the number of observed cases in order to tune the posterior distribution of the model.  *Covariance Matrix:* A matrix showing how all variables interact with one another. Naive Bayes assumed there is no variance between input variables.  *Discretization:* Creates a qualitative variable/ class from a quantitative/ continuous variable X to avoid underlying distribution of the quantitative variable |
| 1. **Random Forest**   *Model Definition:* A decision tree is a classification hierarchical structure that break down complex decision-making process by splitting a number of informative features to make classification. Random Forest is built on a collection of uncorrelated decision trees where the prediction by random features is more accurate than an individual tree. The classification is made by the majority class voted by all trees.[[1]](#footnote-1)  *Split Criterion*: Determine the formula on how to split branch code in tree models  *MaxNumSplits:* Maximal number of decision splits/ branch nodes of the tree model  *NumVariablesToSample:* Determine the number of random features to be considered in node splitting process  *Pruning:* A technique to reduces the size of tree models by removing some parts/ branches that do not provide power for classification and to prevent overfitting  *MinLeafSize:* Minimum number of observations in each leaf node, which determine number of noises accepted by the model |

1. L. Brieman, “Random Forest” in Machine Learning, vol. 45, no. 1, pp. 5-32, 2001 [↑](#footnote-ref-1)