**Missing Value Imputation**

To handle the missing training data, there are two categories of techniques exists, model based and non-model based approaches. Non-model based techniques includes mean imputation and hot-deck imputation. These techniques generally decrease the variance estimates in statistical procedures. Furthermore, these techniques also results in standard errors and bias in results. On the other hand, model based approaches includes data mining algorithm techniques to predict the missing values. (For ex - Regression model, decision tree, NB etc). This approach results decreasing the variance as well as bias.

For our project, we used three methods to impute the missing data which includes model based and non-model based techniques.

1. Simple Mode Data :- This is the example of a non-model based approaches. According to this approach, we fill the missing feature values with the most frequent data for the respective feature.

2. KNN Based Imputation :- This is the example of a model based approaches. K nearest neighbour method is based on the given training observations, an aggregation of the k values of the nearest neighbours is used as the imputes values. ( In this method, the aggregation type depends on the type of variable).

For our implementation, We used a built-in package of R (VIM) for KNN based imputation. In VIM package, the distance computation for defining the nearest neighbour is based on an extension of the growers distance algorithm, which can handle distance variable of type binary, categorical, ordered, continuous and semi-continuous. In our project, all the missing value features were categorical type, therefore we used growers distance function for to compute the distance between the observations.

Under this approach, there are two methods to impute the missing data using k nearest neighbour observations. First method of aggregation is to use the category with the most occurrences in the k nearest neighbours, if this results in a tie, a category from the tied categories will be randomly chosen. Second method of aggregation is to sample the category from the categories in the k nearest neighbours with probabilities equal to the occurrences in the k values.

For our implementation, we used the first method is aggregation to impute the missing values with the K value as "5". We also tried the imputation with different values of K as 3,5,7,10 and 15 while measuring the model performance using logistic regressions classifier with cross validation. However, there were no significant gain by increasing the K value. Therefore we fix our KNN imputation model with K valued as "5".

<https://cran.r-project.org/web/packages/VIM/VIM.pdf>

3. Random-Forest Based Imputation :- This is the another example of a model based approaches where multiple D-trees are built which contains information corresponding to attributes in the given dataset. This information is used to follow a given set of input attribute, depending on categorical nodes. After which, "Random forest" algorithm is used to generalize ensembles of D-trees through bagging which combines multiple random D-trees to aggregate the prediction for missing data.

For our implementation, We used a built-in package of R (missForest) for Random Forest based imputation. In missForest package, for each feature (Having missing values), it builds the random forest based on the given observations and then predicts the missing values. The algorithm continues to repeat these two steps until stopping criteria is met or the user specific maximum of iterations is reached.

According to default stopping criteria, After each iteration the difference between the previous and the new imputed data matrix is assessed for the continuous and categorical features. The default stopping criteria is defined such that the imputation process is stopped as soon as both differences have become larger once. In case of only one type of variable the computation stops as soon as the corresponding difference goes up for the first time. However, the imputation last performed where both differences went up is generally less accurate than the previous one. Therefore, whenever the computation stops due to the stopping criterion the before last imputation matrix is returned. In our case, we only had the missing values in categorical features, therefore second procedures applies.

Additionally, missForest algorithm provides the user with an estimate of the imputation error. This estimation is based on out-of-bag (OOB) error estimate of random forest.

<https://stat.ethz.ch/education/semesters/ss2013/ams/paper/missForest_1.2.pdf>