REPORT

(Submitted by Mohit Agarwala-19307R004)

1. OBSERVATIONS YOU MADE OF THE DATASET PROVIDED FOR THIS COMPETION

Data Set provided were:

1. train.csv – To train & validate the model. This is a labeled data(size = 1028 x 34)

2.test.csv – To find the accuracy of the model on unlabelled data. This is an unlabeled data (size = 1028 x 33)

Training data: 33 out of 34 columns were the features to train the model and the 2nd column i.e., attrition was the output (labeled) data, based on which, the model optimized its prediction (using loss function).

This data set presents an employee survey, indicating if there is attrition or not. The data set contains 1028 entries. Given the limited size of the data set, the model should only be expected to provide modest improvement in identification of attrition vs.a random allocation of probability of attrition.

It has gathered information on employee satisfaction, income, seniority and some demographics.

2. WHAT ALL PREPROCESSING METHODS YOU USED AND WHY?

The given Dataset consists of features that are responsible for Employee Attrition in a Company. The task at hand was to develop a machine learning model that could predict the Attrition of a employee given a set of features with maximum accuracy.

The steps taken are the following :

• Data Exploration

• Data Pre-Processing

• EDA

• Dropping of irrelevant data

• Fit models and Predict Data

Data Pre-processing is necessary as in most datasets there maybe missing data, data containing various degrees of noise and the feature data are not scaled. The following are the techniques used for pre-processing of data:

• **Missing value Imputation and Duplicate Removal** : In this case, the dataset is devoid of any missing values and duplicates. In general , for ”numerical” features the missing values are replaced with their MEDIAN and for ”categorical” features the MODE is opted.

• **Encoding Categorical Data** : Most ML Algorithms like Logistic Regression, SVM, etc. cannot work with Categorical data. So, it is imperative that these features be numerically encoded by either One-

Hot Encoding or Label Encoding. In the following implementation, Label Encoding is used as One-Hot Encoding increases the dimensionality of the dataset. Scikit-learn modules have been used for encoding.

• **Feature Selection :** Feature selection methods are used to improve the classifier's predictive ability by reducing the feature dimensions. The irrelevant features of the Dataset are dropped and then used for fitting the models and predicting the outcomes.

• **Feature Scaling** : This plays an important part for any ML model. Normalization or Feature Scaling is used to reduce the effect of overshooting of weights for some parameters and biasing, and to achieve good results

3. LIST OF VARIOUS APPROACHES YOU USED FROM THE START OF THE COMPETITION TILL YOUR FINAL APPROACH FOR BEST ACCURACY

In order to decide which algorithms would be best for the highest accuracy, the following models were tested and their accuracy score were recorded.

• **kNN classifier** : In k-nearest neighbors (k-NN) classifier a data vector is classified by a plurality vote by its neighbors, with the vector being assigned the class that gets the most votes among its k-neighborhood. It is sensitive to local structure of the data.

• **Gaussian Naive Bayes** : Naive-Bayes algorithms work on the principle of Bayesian Thorem , i.e. conditional probability. Gaussian Naive Bayes (GNB) algorithm is a special type of Naive-Bayes algorithm. It's specifically used when the features have continuous values. It's also assumed that all the features are following a gaussian distribution i.e, normal distribution.

• **Multi-layer Perceptron (MLP) Classifier** : Unlike other classification algorithms such as Support Vectors or Naive Bayes Classifier, MLPClassifier relies on an underlying Neural Network to perform the task of classification. This model optimizes the log-loss function using LBFGS (limited memory broyden fletcher goldfarb shanno) algorithm.

• **Logistic Regression** : The Logistic Regression model is built on the basic model of linear regression with th difference being that, it is used for classification problems and hence uses cross entropy as loss function and as a non-linear activation function uses sigmoid function. It is a form of binary regression.

• **Support Vector Classifier** : For a dataset consisting of features set and labels set, an SVM classifier builds a model to predict classes for new examples. It assigns new example/data points to one of the classes. If there are only 2 classes then it can be called as a Binary SVM Classifier. Here a non-linear SVM Classifer is used with the Radial Basis Function kernel. $K(x,x\,0) = e\,{-kx}\,{-}\,x\,0\,k\,2\,2\sigma\,2$ where x,x 0 are vectors of the feature space, σ is a hyperparameter.

• **Random Forest Classifier** : A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The hyperparameters for this model are the following : n estimators(the number of trees in the forest),criterion(measures split quality). Each tree in the forest

gives a classification, and we say the tree ”votes” for that class. The forest chooses the classification having the most votes (over all the trees in the forest).

XGBClassifier:

The XGBoost stands for Extreme Gradient Boosting and it is a boosting algorithm based on Gradient Boosting Machines. XGboost applies regularization technique to reduce overfitting, and it is one of the differences from the gradient boosting. Another advantage of XGBoost over classical gradient boosting is that it is fast in execution speed.

This gave me the highest accuracy over test data.

4.RESULTS AND FINAL LEARNING YOU ACHIEVED THROUGH THIS COMPETITION

I have learned many python and data science libraries such as numpy,pandas,scikit learn,etc.

I have learnt how to get clean and preprocess raw data which is a big step in ML and data field.

I have learnt to use kaggle ,which I never used before.

In order to complete the assignment,I had to went through many errors and some of them took time to get solved but finally I got through all of that.