Sameer Kishor Kulthe

Evaluation Phase Project Article

**TITANIC PROJECT**

**INTRODUCTION**

* The problem is based on Titanic the unsinkable ship in early 1912. Various attributes are given based on which we need to build a model which would predict whether the arbitrary passenger on Titanic would survive or not.
* The reason behind choosing this problem is that we could build a model to predict the survival rate of passengers and take any proactive decisions to either stop or be better prepared if an event like this were to occur in the future. This would help us to reduce the fatality rate in the future by taking appropriate measures as well.

**Data Analysis**

* Since the target variable which is “Survived” is either a “yes” or a “no”, it is a classification problem
* Call all the required libraries in the beginning which would help us to perform the analysis and facilitate the model building process.
* The dataset is available is a csv file format. Hence, we load the dataset first in the csv file format.
* We then convert the csv file into a dataframe.
* Once we have converted the csv file into a dataframe, we write a code to list the columns, find the shape of the data set and check for any null values in the dataset if any. The dataset has integer, float and object type variables.
* The following columns have been removed as they seem to have no significant contribution to the outcome which are: PassegnerId,Name,Ticket,Cabin and Embarked.
* We came across 177 null values in the age attribute. We then fill the null values with the mean values of the age attribute.
* We check again if the missing values have been filled or not. In our case, we find that the missing values have been filled.
* Label Encoder is used to convert the object data type to numeric one for further use.
* A code is written to describe the dataset whereby, we find the details like count, mean, standard deviation, minimum value, 25th percentile,50th percentile, 75th percentile and the maximum value which helps one to identify certain insights like the presence of outliers or not, skewness and spread of the data from the mean in terms of standard deviation.
* Univariate analysis is performed where we use matplot library and seaborn library are used. A box plot and distplot is used to check for possible outliers, skenweness and countplot.

Chart, bar chart

Description automatically generated

This tells us about the count of attribute sex

Chart, box and whisker chart

Description automatically generated

This plot tells us about the outliers which are present and can be treated if the data loss is within the range of 7-8% of data loss.

**Chart, histogram

Description automatically generated**

This plot tells us that skewness is present to the right

* Bivariate analysis is performed where we use matplot library and seaborn library are used. A bar plots and swarmplots are used to check for the relationships between the dependent and the independent variables.

Chart, bar chart

Description automatically generated

This graph tells us about the relationship of the survival rate with the Pclass. If Pclass 1, the survival rate is high.

Chart, scatter chart

Description automatically generated

This plot tells us about the density of survival or not lies between which age groups the most.

* We have some outliers which we remove it using zscore. Since the loss of data is within the acceptable range, we persist with the new dataset obtained after the removal of outliers.
* Now, we check for the skewness. Attributes like the Pclass, SibSp,Parch and Age have some skewness since the values are greater than 0.5. Power transform function is used to remove the skewness.
* Once we use that function, the skewness is removed.

This concludes the EDA analysis and the dataset is now ready to be used for Machine Learning

**Building Machine Learning Models**

* Now the dataset is ready to be fed to the algorithm to build a model.
* However before using the dataset for model building, the x and y i.e. the dependent and independent variables are separated.
* Standard scaler is used on the x variables to standardise the independent variables. This is done to scale down the values so there is no biasness during the machine learning phase.
* Once the above steps are done, train\_test\_split is used by defining a random\_state and test\_size.
* In order, to find the best fit model, we use four algorithms to decide the best fit model amogst the four.
* After using Logistics Regression, Knn(Nearest Neighbors Classifiers), RandomForest Classifier and Adaboost Classifier, we find that Adaboost is the best fit model as the difference between the accuracy score and the cross validation score is the least in that case.
* Also confusion matrix and classification matrix is printed to get the true positive, true negative, false positive and false negative and the precision, recall, f1-score and support.
* Hyper tuning of parameters is performed by using RandomizedSearchCV. Best estimators, best parameters and best score are found.
* The best best obtained parameters are fed to the adaboost model and the best accuracy score is obtained.
* The plot of auc\_roc\_curve is made to find the relationship between true positive and false positive rate. This plot is nothing but the graphical representation of the confusion matrix
* Finally the model is dumped by performing serialisation for further use by creating an obj file. This file can be deployed if and when needed for further use.

**CONCLUSION**

The accuracy score of the model obtained after hyper tuning the parameter is 0.835 which seems to be a fairly high. True positive is 94, true negative is 43, False positive is 10 and false negative is 17 as per the confusion matrix.

**Global Power Plant Database**

**INTRODUCTION**

* The problem is based on Global Power Plant Database to make predictions on the primary fuel used and the capacity in MegaWatts. Various attributes are given to predict both the target variables
* The reason behind choosing this problem is to make a prediction on the primary fuel and capacity needed which would help us to predict the power consumption and which fuel is used primarily and make us future ready to take required actions to meet the energy requirements.

**Data Analysis**

* We have two target variables i.e. primary fuel and capacity. We will be building two separate models for the two different target variables. The first problem to predict the primary fuel is a classification problem and the second problem to identify the capacity is a regression problem
* Call all the required libraries in the beginning which would help us to perform the analysis and facilitate the model building process. Let’s first address the first part of the problem i.e.the primary fuel prediction.
* The dataset is available is a csv file format. Hence, we load the dataset first in the csv file format.
* We then convert the csv file into a dataframe.
* Once we have converted the csv file into a dataframe, we write a code to list the columns, find the shape of the data set and check for any null values in the dataset if any. The dataset has float and object type variables.
* The following columns have been removed as they seem to have no significant contribution to the outcome or a big chuck of data is missing for which no missing value treatment would hold good either. These columns are:

country,country\_long,name,gppd\_idnr,other\_fuel1,other\_fuel2,other\_fuel3,generation\_gwh,commissioning\_year,owner,url,wepp\_id,year\_of\_capacity\_data,generation\_data\_source,estimated\_generation\_gwh.

* We came across some null values in the following attributes which are latitude, longitude, geolocation\_source, generation\_gwh\_2014, generation\_gwh\_2015, generation\_gwh\_2016 and generation\_gwh\_2017. Based on the attributes we fill the missing values either by mean, median or mode. We have replaced latitude and longitude with mean, geolocation\_source with mode, generation\_gwh\_2014, generation\_gwh\_2015 generation\_gwh\_2016 and generation\_gwh\_2017 with mode.
* We check again if the missing values have been filled or not. In our case, we find that the missing values have been filled.
* A code is written to describe the dataset whereby, we find the details like count, mean, standard deviation, minimum value, 25th percentile,50th percentile, 75th percentile and the maximum value which helps one to identify certain insights like the presence of outliers or not, skewness and spread of the data from the mean in terms of standard deviation.
* Label Encoder is used to convert the object data type to numeric one for further use.
* Univariate analysis is performed where we use matplot library and seaborn library are used. A box plot and distplot is used to check for possible outliers, skenweness and countplot..

Chart, bar chart

Description automatically generated

Since the target variable is imbalanced, we will be using SMOTE to balance the target variable.

Chart, histogram

Description automatically generated

Skewness is present to the right

Chart, box and whisker chart

Description automatically generated

No outliers are present. minimum value is 8.1. Median value is 21.2 and maximum value is 34.6

Chart, histogram

Description automatically generated

Maximum density is between 72.5 and 80. Skewness is present to the right.

Chart, scatter chart

Description automatically generated

Coal is the most used primary fuel having highest capacity.

Chart, scatter chart

Description automatically generated

The range for primary fuel 3 is used the most across latitude.

Chart

Description automatically generated

The range for primary fuel 1 is used the most across longitude.

Chart, scatter chart

Description automatically generated

Generation in terms of gwh is the most with coal in the year 2014,2015,2016 and 2017

* Now, we check for the skewness. Attributes like the generation\_gwh\_2014/2015/2016/2017 have some skewness since the values are greater than 0.5. Power transform function is used to remove the skewness.
* Once we use that function, the skewness is removed.

This concludes the EDA analysis and the dataset is now ready to be used for Machine Learning

**Building Machine Learning Models**

* Now the dataset is ready to be fed to the algorithm to build a model.
* However before using the dataset for model building, the x and y i.e. the dependent and independent variables are separated.
* Standard scaler is used on the x variables to standardise the independent variables. This is done to scale down the values so there is no biasness during the machine learning phase.
* We use SMOTE to balance the target variable.
* Once the above steps are done, train\_test\_split is used by defining a random\_state and test\_size.
* Inorder, to find the best fit model, we use four algorithms to decide the best fit model amogst the four.
* After using Support Vector Classification, Knn(Nearest Neighbors Classifiers), RandomForest Classifier and Adaboost Classifier, we find that RandonForest is the best fit model as the difference between the accuracy score and the cross validation score is the least in that case.
* Also confusion matrix and classification matrix is printed to get the true positive, true negative, false positive and false negative and the precision, recall, f1-score and support.
* Hypertuning of parameters is performed by using RandomizedSearchCV. Best estimators, best parameters and best score are found.
* The best best obtained parameters are fed to the randomforest model and the best accuracy score is obtained.
* Finally the model is dumped by performing serialisation for further use by creating an obj file. This file can be deployed if and when needed for further use.
* Same steps are followed for the target variable with capacity. Just that this is a regression problem so the algorithms used are Linear Regression, KNN Regressor, Lasso and Ridge.
* For the second case, RandomForest Regressor is the best fit model.
* Rest of the steps are followed like for the classification problem.

**CONCLUSION**

The accuracy score of the first model obtained after hyper tuning the parameter is 0.91 which seems to be a fairly high. The accuracy score for the second model after hyper tuning is 0.90 which also seems to be a fairly high value. The primary fuel used is coal as per the study.