Articles/Blogs

I have used two projects in articles/blogs. Those who have been used in evaluation practice phase. Which is as follows:–

* Insurance Fraud Detection
* Flight Price Prediction Training and Testing

These sub - topics are used in articles:-

* Problem Definition
* EDA Concluding Remarks
* Pre-Processing Pipelines
* Building Machine Learning Models
* Conclusion Remarks

Now I will discuss both one by one projects. Firstly I will discuss the Insurance Fraud Detection project.

Discuss the project insurance fraud detection:-

* Problem Definition:-

This project has a logistic classification problem because the output in it is in the form of yes or no. This mean it has fraud customer or not.

Multiple columns and rows are used in this project which shape is (1000, 40) . These include the output column fraud reported.

We will create the best model using the input/output variables in this dataset then by using the best model , we will be able to perform better prediction in the new input data.

Now next I will use EDA process .

* EDA Concluding Remarks:-

In this , I have used multiple methods , attributes so that we can understand the data well such as –

* Shape – check that how many row and columns are there in this dataset.
* Columns – checking the which columns are used in this dataset.
* Info() – what does this method describe that how many count values , particular data types , null values , columns name in column wise.
* Is null().sum() – check the null/missing values in dataset and there is no null values in this dataset.
* Describe() – this method gives a column wise statistical value.
* Corr() – this method describe that the input and output variables and how many are correlated to each other?
* Skew() – this method show that skewness present or not in dataset and skewness always show in continuous values and there is no skewness value I this dataset and thus we will not use any skewed removal methods in this.
* Check the outliers – to check the outliers best methods is box plot . Zscore method used in remove outliers. In this dataset there is no outliers present and only this dataset is biased.
* Plot() – this method used in box and density types all the columns. This is the method visualize method.

I have used in multiple visualize methods which is as follows:-

* Countplot() – this visualize methods only use univariate variable(categorical).
* Violinplot() - this visualize methods also use univariate variable(continuous).
* Scatterplot() - this visualize methods use bivariate variable(continuous).
* Histplot() – this visuailize methods also use univariate variable(continuous).
* Heatmap() - this visuailize methods also use multivariate variable(continuous).The heatmap show that correlated data.
* Distplot() - this visuailize methods also use univariate variable(continuous).This distribution plot show that easily understand skewness in input variables.
* Pre-Processing Pipeline:-

In this project multiple pipeline used which is as follows –

* Ordinal Encoding – this methods change the categorical value to continuous value .
* Min Max Scaling – this method use in dataset all value set in same range(0 to 1).
* Grid Search CV – this methods tuning of the particular best model choose.
* Cross-Validation – this method resampling procedure used to evaluate model.

This project multiple model used for the better prediction which is as follows –

Logistic Regression , Decision Tree Classifier , SVC , AdaBoost Classifier , GaussianNB , Random Forest Classifier , KNeighbours Classifier.

* Building Machine Learning Models:-

In this project I have used different models which are written above but before using these models, I have to use best random state and I will use a logistic model for the best random state.

Me on using the logistic model then I found that best random state value 97 and accuracy score 80%.Now I will use the random state value 97 in all models.

I found out the accuracy score all models which is as follows –

Model Name Accuracy Score

Logistic Regression - 78.69%

Decision Tree Classifier - 79.13%

SVC - 74.78%

Ada Boost Classifier - 80.86%

Gaussian NB - 72.17%

K Neighbors Classifier - 72.60%

Random Forest Classifier - 79.56%

These models accuracy score is data biased so that remove biasness in this dataset , I will use cross validation technique and I will find out the cross val score in these models which is as follows –

Model Name Cross Val Score

Logistic Regression - 76.70%

Decision Tree Classifier - 77.89%

SVC - 74.89%

Ada Boost Classifier - 79.39%

Gaussian NB - 71.60%

KNeighbors Classifier - 71.89%

Random Forest Classifier - 77.40%

* Conclusion Remark:-

Now further to choose best model , I need to see of these models minimum difference accuracy score and cross val score and the model that will have the least difference , The same model would be best for this dataset.

Minimum difference is accuracy and cross val score SVC() model so this is our best model.

Now further I will use Grid Search CV technique for the best parameter choose in the model and In this technique , my best parameters are the kernel value ‘linear’ and C value 10 .

The model accuracy score is increased when Use this parameter in SVC() model .

Now I will import Joblib Library and I will save this model which further became useful in using prediction with new input data .I will load the model from the library to do prediction and I will able to save that prediction.

Discuss The Project Flight Price Prediction Training and Testing :-

* Problem Definition:-

I will divide flight price prediction project into two phase, one is training and other one is testing to dataset so firstly I will use flight price prediction training phase.

This project has a linear Regression problem because the output in it is in the form of integer value.

Multiple columns and rows are used in this project which shape is (10683, 11) . These include the output column Price.

In this using multiple columns such that particular airline, date of journey, source, destination, route, duration, total stops etc ,the price of the particular passenger has been set value.

I will create the best model using the input/output variables in this dataset then by using the best model , I will be able to perform better prediction in the new input data.

Now next I will use EDA process .

* EDA Concluding Remarks:-

In this , I have used multiple methods , attributes so that we can understand the data well such as –

* Shape – check that how many row and columns are there in this dataset. This dataset input columns 10 and output column is 1 and 10683 rows use.
* Columns – checking the which columns are used in this dataset. In this dataset 11 columns are used.
* Info() – what does this method describe that how many count values , particular data types , null values , columns name in column wise.
* Is null().sum() – check the null/missing values in dataset and there is two null values in this dataset. One null value is route column and other one null value total stops present, these null value remove I will use simple imputer method because I have to use strategy ‘most frequent’ to remove object type all null value in this method.
* Describe() – this method gives a column wise statistical value such that mean, median, maximum value, minimum value, standard deviation, 25%,75%,50% percentile etc. I have used heatmp to visualize the describe method.
* Corr() – this method describe that the input and output variables and how many are correlated to each other? I have used heatmap to visualize the correlation method and In this dataset price column highly correlated total stops column and very less correlated depature time .
* Skew() – this method show that skewness present or not in dataset and skewness always show in continuous values and there is no skewness value because all input variables present in categorical(object) form thus I will not use any skewed removal methods in this dataset.
* Check the outliers – to check the outliers best methods is box plot . Zscore method used in remove outliers. In this dataset there is no outliers present and only this dataset is biased and all input variables datatype object type and output variable is continuous but this is predict column so that no need remove the outliers.
* Plot() – this method used in box and density types all the columns . This is the method visualize method.

I have used in multiple visualize methods which is as follows:-

* Countplot() – this visualize methods only use univariate variable(categorical).
* Violinplot() - this visualize methods also use univariate variable(continuous).
* Scatterplot() - this visualize methods use bivariate variable(continuous).
* Histplot() – this visuailize methods also use univariate variable(continuous).
* Heatmap() - this visuailize methods also use multivariate variable(continuous).The heatmap show that correlated data.
* Distplot() - this visuailize methods also use univariate variable(continuous).This distribution plot show that easily understand skewness in input variables.
* Pre- Processing Pipeline:-

In this project multiple pipeline used which is as follows –

* Ordinal Encoding – this methods change the categorical value to continuous value and It can be easily understand machine model.
* Min Max Scaling – this method use in dataset all value set in same range(0 to 1) which will model score very well.
* Grid Search CV – this methods tuning of the particular best model choose. This technique use for increase the model accuracy score.
* Cross-Validation – this method resampling procedure used to evaluate model. This technique used after checking the different models so that I can make the best model choose for this dataset.

This project multiple model used for the better prediction which is as follows –

Linear Regression, Elastic Net Regressor, Random Forest Regressor, Ada Boost Regressor, Decision Tree Regressor.

* Building Machine Learning Model:-

In this project I have used different models which are written above but before using these models, I have to use best random state and I will use a linear model for the best random state.

Me on using the linear model then I found that best random state value 192 and r2 score 39.80%.Now I will use the random state value 192 in all models.

I found out the r2 score all models which is as follows –

Model Name Accuracy Score

Linear Regression - 39.80%

Decision Tree Regressor - 80.93%

Elastic Net - 39.80%

Ada Boost Regressor - 71.11%

Random Forest Regressor - 89.37%

These models r2 score is in data biased so that remove biasness in this dataset , I will use cross validation technique and I will find out the cross val score in these models which is as follows –

Model Name Cross Val Score

Linear Regression - 35.26%

Decision Tree Regressor - 74.94%

Elastic Net - 35.27%

Ada Boost Regressor - (-5.06%)

Random Forest Regressor - 86.01%

* Conclusion Remarks:-

Now further to choose best model , I need to see of these models minimum difference accuracy score and cross val score and the model that will have the least difference , The same model would be best for this dataset.

Minimum difference is accuracy and cross val score Random Forest Regressor() model so this is our best model.

Now I will import Joblib Library and I will save this model which further became useful in using prediction with new input data .I will load the model from the library to do prediction and I will able to save that prediction.

Flight Price Prediction for Testing:-

* Problem Definition:-

Multiple columns and rows are used in this project which shape is (2671, 10) . This dataset no output column present because this data only use the input data for prediction. It has to be prediction that by using the passenger detail and how much does that passenger have to pay price for the particular airline.

In this using multiple columns such that particular airline, date of journey, source, destination, route, duration, total stops etc .

Now next I will use EDA process .

* EDA Concluding Remarks:-

In this , I have used multiple methods , attributes so that we can understand the data well such as –

* Shape – check that how many row and columns are there in this dataset. This dataset only input columns 10 and no output column and 2671 rows use.
* Columns – checking the which columns are used in this dataset. In this dataset 10 columns are used.
* Info() – what does this method describe that how many count values , particular data types , null values , columns name in column wise.
* Is null().sum() – check the null/missing values in dataset and there is no null values in this dataset thus no need null value removal method.
* Describe() – this method gives a column wise statistical value such that mean, median, maximum value, minimum value, standard deviation, 25%,75%,50% percentile etc.
* Corr() - this method describe that the input and output variables and how many are correlated to each other? I have used heatmap to visualize the correlation method but correlation only input variables .
* Skew() – this method show that skewness present or not in dataset and skewness always show in continuous values and there is no skewness value because all input variables present in categorical(object) form thus I will not use any skewed removal methods in this dataset.
* Check the outliers – to check the outliers best methods is box plot . Zscore method used in remove outliers. In this dataset there is no outliers present and only this dataset is biased and all input variables datatype object type so that no need remove the outliers.
* Plot() – this method used in density types all the columns . This is the method visualize method.

I have used in multiple visualize methods which is as follows:-

* Countplot() – this visualize methods only use univariate variable(categorical).
* Stripplotplot() - this visualize methods also use bivariate variable(continuous).
* Scatterplot() - this visualize methods use bivariate variable(continuous).
* Histplot() – this visuailize methods also use univariate variable(continuous).
* Heatmap() - this visuailize methods also use multivariate variable(continuous).The heatmap show that correlated data.
* Distplot() - this visuailize methods also use univariate variable(continuous).This distribution plot show that easily understand skewness in input variables.
* Final Prediction For Testing:-

I have already trained the model for this project. I have to load the model to predict it now. For this I will load the model from the joblib library and predict test data in that model.

I can predict the flight price of any passenger.

For example :-

I used the details of a passenger such as airline-indigo, date of journey- 12/05/2019, source-kolkata, destination- banglore, dep\_time- 06:20,arrival\_time-10:20, duration-4hour, total\_stops-1stop.After using these details the passenger will have to pay 7065.57 this price.