NAME OF THE PROJECT

FLIGHT PRICE PREDICTION PROJECT

Submitted by: **AKSHAY RUNTHALA**

ACKNOWLEDGMENT

I took the help of various sources (mentioned below) to complete my project:-

- **SCIKIT-LEARN DOCUMENTATION(https://scikit-learn.org/stable/)**
- **NOTES PREPARED BY MYSELF AND RESOURCES PROVIDED BY MY TRAINING INSTITUTE(DATA TRAINED)**
- **PANDAS DOCUMENTATION(https://pandas.pydata.org/docs/)**
- **MATPLOTLIB DOCUMENTATION(https://matplotlib.org/)**
- **SEABORN DOCUMENTATION(https://seaborn.pydata.org/)**

INTRODUCTION

- **Business Problem Framing**
  Describe the business problem and how this problem can be related to the real world.

MY ANSWER-
The business problem I came across while preparing this project was that if any business/individual chooses to do business in flight booking,then it is very important to find out the value of the flights as accurately as possible.The flights' price change from one company to another and also varies with time and the customers/travelers consider price to be the main factor while booking the flights.**If the value of flight is fixed at very high rate, then customers will avoid that flight unless he/she needs to travel due to emergency**.

I can say the accuracy of estimate and time taken to estimate the price/value of a flight are the main limitations faced in today's world nowadays.The values predicted by the flight booking agent in the end are an estimate of a human and actual price of the used flight varies from estimate.Many businesses/individuals rely on the estimates/decision of the flight estimate/agents to book a particular flight for the particular date.

There is a need for a pattern to estimate the flights' values as accurately as possible. Human beings' abilities are limited due to time factor as well as accuracy in estimation of values.
Machine learning tasks can easily overcome these limitations of human beings because they are exponentially faster than human estimators to predict the value of a flight based on the given data (flight features).Most importantly,the machine learning models can predict the values of flights more accurately than humans.

Conceptual Background of the Domain Problem

Describe the domain related concepts that you think will be useful for better understanding of the project.

MY ANSWER-

- Time and date of the flight journey
- Source and Destination cities of the flight.

Review of Literature

This is a comprehensive summary of the research done on the topic. The review should enumerate, describe, summarize, evaluate and clarify the research done.

MY ANSWER:--

I searched the websites for understanding the meaning of the terms and concepts used in the flight booking industry.I needed to understand these concepts so that I can perform my machine learning tasks better.

● Motivation for the Problem Undertaken

Describe your objective behind making this project, this domain and what is the motivation behind.

MY ANSWER:--

My objective for making this project is to-

- Make a machine learning model that can most accurately predict the **PRICES** of every flight regarding which data is available to the company.
- To make this model as efficient as possible to predict the future flights based on data available to the company.
- Find out the **important variables** affecting the price of a single flight.
- Develop my skills for data science further by undertaking this project. This is a great project which directly relates to the real world scenario. I was motivated by the various ideas that came when I first looked into this project.
- Test and upgrade my various python coding skills while working on my project.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Describe the mathematical, statistical and analytics modelling done during this project along with the proper justification.

I calculated the skewness of the whole dataset to know whether the data is normal or not. I also used skewness to impute missing values in numerical columns of the dataset. If my data is not normal,It is better to impute missing numerical values with the median value and if it is normal,then it is advisable to impute missing values with mean.

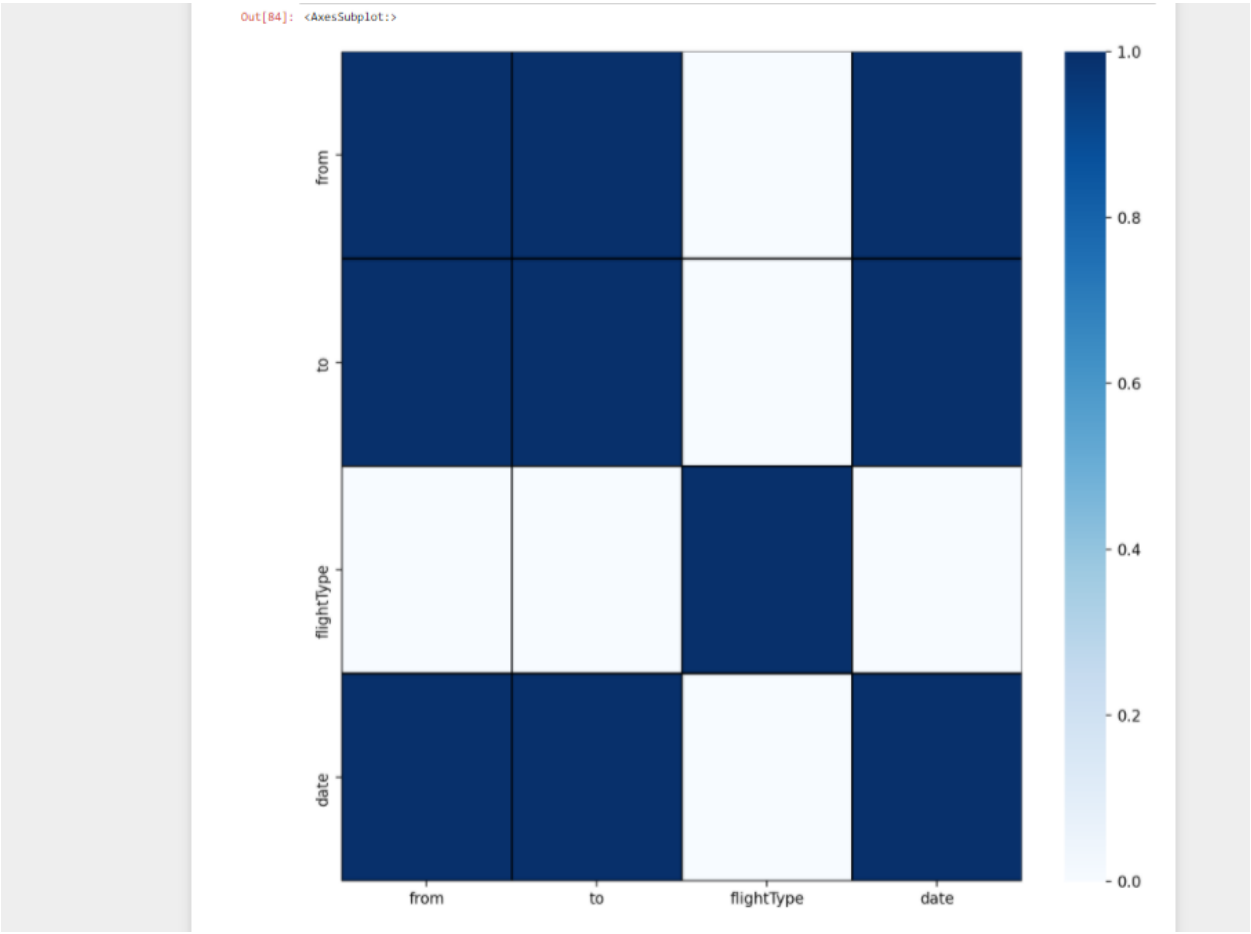
I imputed my categorical columns' missing values with the most frequent value.

SKEWNESS CHECK AND OBSERVATIONS:-

I Plotted Skewness For All Data Distributions Against Columns Of The Data. Plotted Green Horizontal Line For A Skewness Of Zero(Normal Distribution Has A Skewness Of 0). Plotted Red Lines Around The Green Line Denoting A Range Of (- 0.5,0.5).If Data Points Fall Within The Above Mentioned Range,Then They Are Approximately Normally Distributed . The rule for skewness seems to be: If the skewness is between -0.5 and 0.5, the data are fairly symmetrical. If the skewness is between -1 and -0.5 (negatively skewed) or between 0.5 and 1 (positively skewed), the data are moderately skewed. Skewness of the normal distribution is zero.**They all are more or less approximately normal**.**I did not consider skewness for the categorical columns as well as the target label - price in my dataset.** Correlation coefficients are indicators of the strength of the linear relationship between two different variables, x and y. A linear correlation coefficient that is greater than zero indicates a positive relationship. A value that is less than zero signifies a negative relationship. Finally, a value of zero indicates no relationship between the two variables x and y.

I calculated the correlation between my target label (PRICE) and other columns of the data to know which variables affect my price the most.

I used chi square tests from scipy library to find relationships between the categorical variables of the data. Below is the heatmap for chi-square test results.



As shown in the above heatmap figure,the **'blue'** colour shows the relationship status to be **'yes' or (1)**,whereas the **'white'** colour shows the relationship status to be **'no' or (0)** between the variables of the dataset.

- Data Sources and their formats

  What are the data sources, their origins, their formats and other details that you find necessary? They can be described here. Provide a proper data description. You can also add a snapshot of the data.

- ☐ I downloaded a dataset regarding flight data containing 271888 rows of data.
- ☐ I found various important details regarding the flight.They are mentioned below-
    - Destination city of flight
    - Distance covered by the flight
    - Date of journey
    - Source city of flight
    - Time of flight journey

Data Preprocessing Done

What were the steps followed for the cleaning of the data? What were the assumptions done and what were the next actions steps over that?

My Answer--

- **There were no missing values in the dataset.**
- I used the Label Encoding method from sklearn library to convert the ‘object’ data type columns in the data to float or integer.
- I used the Elliptic Envelope to detect the outliers and removed them(the removal was limited to 5% of the dataset).
- **There was no need for adjusting skewness of the dataset because it was approximately normal.**
- I used the Standard Scalar from sklearn.preprocessing to scale the data because scaling was necessary for better regression tasks.

- Data Inputs- Logic- Output Relationships

Describe the relationship behind the data input, its format, the logic in between and the output. Describe how the input affects the output.

My Answer--

The relationship between data input and output is very important in machine learning model preparation. If the data we import is incorrect, then the output will also be incorrect. Every machine learning task is dependent on the data we import to prepare the machine learning model.

I used the correlation matrix to find the relationship between inputs and output. Correlation greater than 0.5 is considered a strong positive relationship.

There was strong positive relationship between every mentioned inputs/columns and the output (Price)

- State the set of assumptions (if any) related to the problem under consideration

My Answer--

1. I chose a random state for all my regression models to be **185.**

2. I chose a random state for splitting my data into training and test data to be **87.**
3. I used an elliptic envelope method to detect the outliers in my dataset **because we can limit the amount of outliers to be removed from the dataset using this method.**

- Hardware and Software Requirements and Tools Used

  MY ANSWER-

  Libraries and packages used :-

  - Pandas-used for Exploratory data analysis

    I used-

    - ☐ .describe() method to understand the data
    - ☐ Pivot table method from pandas library to analyse my data better and derive meaningful insights for my exploratory data analysis.
    - ☐ .info() method to understand the data type of columns in the data.
  - Matplotlib- used for Exploratory data analysis
  - Seaborn- used for Exploratory data analysis
    - ☐ Used bar plot,scatter plot to visualize my data for EXPLORATORY DATA ANALYSIS.
    - ☐ Used **heatmap** to show **correlation/relationship** between the columns/variables of the data.
    - ☐ Used boxplot to detect outliers in my data.
  - Sklearn-used for machine learning

- ☐ Imported various regression models like KNN,DECISION TREE,RANDOM FORESTS,LIGHT xgboost REGRESSOR,XGBOOST
- ☐ Performed standard scaling using standard scaler from preprocessing.

- Scipy-used for statistics
  - ☐ For computing Chi-square tests.**CHI-SQUARE TEST IS USED TO DETERMINE THE RELATIONSHIP BETWEEN THE CATEGORICAL VARIABLES OF THE DATASET.**
- Light gbm regression –used for machine learning
- Xgboost regression -used for machine learning

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  Describe the approaches you followed, both statistical and analytical, for solving this problem.

  MY ANSWER-

  - ☐ I calculated the skewness method using the pandas library and plotted the skewness data to determine the normality of the data because I found it more effective.

- ☐ I used the ***ELLIPTIC ENVELOPE*** method to detect my outliers upto a predefined limit(I chose around 5% of my data as outliers).

- Run and Evaluate selected models

Describe all the algorithms used along with the snapshot of their code and what were the results observed over different evaluation metrics.

MY ANSWER-

Listing down all the algorithms used for the training and testing.

- ☐ RANDOM FOREST REGRESSOR
- ☐ ADABOOST-REGRESSOR
- ☐ XGBOOST REGRESSOR
- ☐ LIGHT GBM REGRESSOR
- ☐ ELASTIC NET REGRESSION
- ☐ K-NEAREST NEIGHBORS REGRESSOR
- ☐ DECISION TREE REGRESSOR
- ☐ SUPPORT VECTOR MACHINE REGRESSION

ElasticNet()

Root_mean_squared_error: 264.29154544665363

mean_absolute_error: 217.34230120254185

mean_squared_error: 69850.02099458058

r2: 0.4526292365940948

cross validation scores below:-- ElasticNet()

root_mean_squared_error: -262.3565384079286

mean_absolute_error_cross_val_score: -217.56570484175376

mean_squared_error_cross_val_score: -69575.37607463646

DecisionTreeRegressor(random_state=185)

Root_mean_squared_error: 36.91112609868618

mean_absolute_error: 24.80093253981126

mean_squared_error: 1362.431229873112

r2: 0.989323481771873

cross validation scores below:-- DecisionTreeRegressor(random_state=185)

root_mean_squared_error: -37.94421969690226

mean_absolute_error_cross_val_score: -25.047118906250933

mean_squared_error_cross_val_score: -1443.8428717634881

RandomForestRegressor(random_state=185)

Root_mean_squared_error: 33.818065122050506

mean_absolute_error: 24.464344907691604

mean_squared_error: 1143.661528599249

r2: 0.9910378425793758

cross validation scores below:-- RandomForestRegressor(random_state=185)

root_mean_squared_error: -34.05202613186458

mean_absolute_error_cross_val_score: -24.5140588226976

mean_squared_error_cross_val_score: -1160.5393667366275

AdaBoostRegressor(random_state=185)

Root_mean_squared_error: 135.26602268701834

mean_absolute_error: 111.8594950200967

mean_squared_error: 18296.89689356496

r2: 0.856618705649826

cross validation scores below:-- AdaBoostRegressor(random_state=185)

root_mean_squared_error: -134.4144448449893

mean_absolute_error_cross_val_score: -111.0381955753248

mean_squared_error_cross_val_score: -18130.647425737927

GradientBoostingRegressor(random_state=185)

Root_mean_squared_error: 49.386046636113434

mean_absolute_error: 37.884739183922804

mean_squared_error: 2438.9816023443714

r2: 0.9808872323501266

cross validation scores below:-- GradientBoostingRegressor(random_state=185)

root_mean_squared_error: -54.29601206768024

mean_absolute_error_cross_val_score: -40.85678496896092

mean_squared_error_cross_val_score: -2955.0464229157365

SVR()

Root_mean_squared_error: 146.4811535331914

mean_absolute_error: 110.05055131521868

mean_squared_error: 21456.72834041439

r2: 0.8318570903107247

cross validation scores below:-- SVR()

root_mean_squared_error: -276.72041889619624

mean_absolute_error_cross_val_score: -228.44434952338256

mean_squared_error_cross_val_score: -77333.52889440865

LGBMRegressor(random_state=185)

Root_mean_squared_error: 30.394731754882137

mean_absolute_error: 24.163101989013143

mean_squared_error: 923.8397184512405

r2: 0.9927604481036221

cross validation scores below:-- LGBMRegressor(random_state=185)

root_mean_squared_error: -30.468812313647845

mean_absolute_error_cross_val_score: -24.267402041795172

mean_squared_error_cross_val_score: -928.8631026754326

KNeighborsRegressor()

Root_mean_squared_error: 36.462986165614055

mean_absolute_error: 25.637915687576626

mean_squared_error: 1329.5493601137618

r2: 0.9895811563422756

cross validation scores below:-- KNeighborsRegressor()

root_mean_squared_error: -140.58182272639064

mean_absolute_error_cross_val_score: -98.54865871087186

mean_squared_error_cross_val_score: -19915.428559292835

XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,

colsample_bynode=1, colsample_bytree=1, gamma=0, gpu_id=-1,

importance_type='gain', interaction_constraints='',

learning_rate=0.300000012, max_delta_step=0, max_depth=6,

min_child_weight=1, missing=nan, monotone_constraints='()',

n_estimators=100, n_jobs=8, num_parallel_tree=1, random_state=185,

reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,

tree_method='exact', validate_parameters=1, verbosity=None)

Root_mean_squared_error: 30.39205013662801

mean_absolute_error: 24.030914579546597

mean_squared_error: 923.6767115073108

r2: 0.9927617254867076

cross validation scores below:-- XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,

colsample_bynode=1, colsample_bytree=1, gamma=0, gpu_id=-1,

importance_type='gain', interaction_constraints='',

learning_rate=0.300000012, max_delta_step=0, max_depth=6,

min_child_weight=1, missing=nan, monotone_constraints='()',

n_estimators=100, n_jobs=8, num_parallel_tree=1, random_state=185,

reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,

tree_method='exact', validate_parameters=1, verbosity=None)

root_mean_squared_error: -30.397542150661003

mean_absolute_error_cross_val_score: -24.048628748593877

mean_squared_error_cross_val_score: -924.5270607521379

I used various metrics to evaluate my model performance and based on them, I selected my models for hyper parameter tuning.

In [110]:
```
gr = GridSearchCV(rfr,param_grid=rfprr)
gl = GridSearchCV(lgr,param_grid=lgpr1)
```

I created grid search cv models for the best models chosen above.

In [120]:
```
print(gr)
gr.fit(sX_train,y_train)
y_pred=gr.predict(sX_test)
print('Mean_absolute_error: ',mean_absolute_error(y_test,y_pred))
print('Mean_squared_error: ',mean_squared_error(y_test,y_pred))
print('Root_mean_squared_error: ',np.sqrt(mean_squared_error(y_test,y_pred)))
print('R2_score: ',r2_score(y_test,y_pred))
print(gr.best_params_)
print('\n')
print(gr.best_estimator_)
```

```
GridSearchCV(estimator=RandomForestRegressor(random_state=185),
             param_grid={'max_depth': [None, 12, 10, 8],
                         'max_features': ['auto', 'log2', 'sqrt'],
                         'min_samples_leaf': [1, 2, 3],
                         'min_samples_split': [1, 2, 3],
                         'n_estimators': [100, 200, 300, 500, 700]})
Mean_absolute_error:  23.97519930251283
Mean_squared_error:  924.8285955220697
Root_mean_squared_error:  30.41099464868043
R2_score:  0.9927526988948249
{'max_depth': 12, 'max_features': 'auto', 'min_samples_leaf': 3, 'min_samples_split': 2, 'n_estimators': 700}


RandomForestRegressor(max_depth=12, min_samples_leaf=3, n_estimators=700,
                      random_state=185)
```

In [121]:
```
print(gl)
gl.fit(sX_train,y_train)
y_pred=gl.predict(sX_test)
print('Mean_absolute_error: ',mean_absolute_error(y_test,y_pred))
print('Mean_squared_error: ',mean_squared_error(y_test,y_pred))
print('Root_mean_squared_error: ',np.sqrt(mean_squared_error(y_test,y_pred)))
print('R2_score: ',r2_score(y_test,y_pred))
print(gl.best_params_)
print('\n')
print(gl.best_estimator_)
```

```
GridSearchCV(estimator=LGBMRegressor(random_state=185),
             param_grid={'learning_rate': [0.1, 0.001, 0.2, 0.3, 0.5, 0.7, 0.9],
                         'max_depth': [6, 7, 8, 10, 12],
                         'n_estimators': [200, 300, 500, 700, 800],
                         'n_jobs': [1], 'reg_alpha': [0, 0.5, 1, 1.2],
                         'reg_lambda': [1.2, 0, 1, 0.5]})
Mean_absolute_error:  24.060016766788547
Mean_squared_error:  921.126552615655
Root_mean_squared_error:  30.350066764599628
R2_score:  0.9927817094809779
{'learning_rate': 0.1, 'max_depth': 10, 'n_estimators': 200, 'n_jobs': 1, 'reg_alpha': 0.5, 'reg_lambda': 1.2}


LGBMRegressor(max_depth=10, n_estimators=200, n_jobs=1, random_state=185,
              reg_alpha=0.5, reg_lambda=1.2)
```

I found the best parameters for **RANDOM FOREST regressor** in grid search cv.

I chose the **LIGHT GBM REGRESSOR and RANDOM FOREST REGRESSOR** as the models for hyperparameter tuning because they

performed the best among all the models based on metrics and scores and also r2 score.

I created grid search cv models for the best models chosen above.

RandomForest regressor performed best among all grid search cv models.

· Key Metrics for success in solving problem under consideration

What were the key metrics used along with justification for using it? You may also include statistical metrics used if any.

My answer-

The metrics used were--

- ☐ R2 score-- the closer it is to 1,the better the model performance is.
- ☐ Mean absolute error
- ☐ Mean squared error
- ☐ Root mean squared error(RMSE)

All the errors should be as minimal as possible.the lesser they are,the better the model performance is.

· Visualizations

Mention all the plots made along with their pictures and what were the inferences and observations obtained from those. Describe them in detail.

If different platforms were used, mention that as well.

My answer-

1. Histogram

   I plotted it to know about the skewness and distribution of the numerical data columns.

In statistics, a positively skewed (or right-skewed) distribution is a type of distribution in which most values are clustered around the left tail of the distribution while the right tail of the distribution is longer.

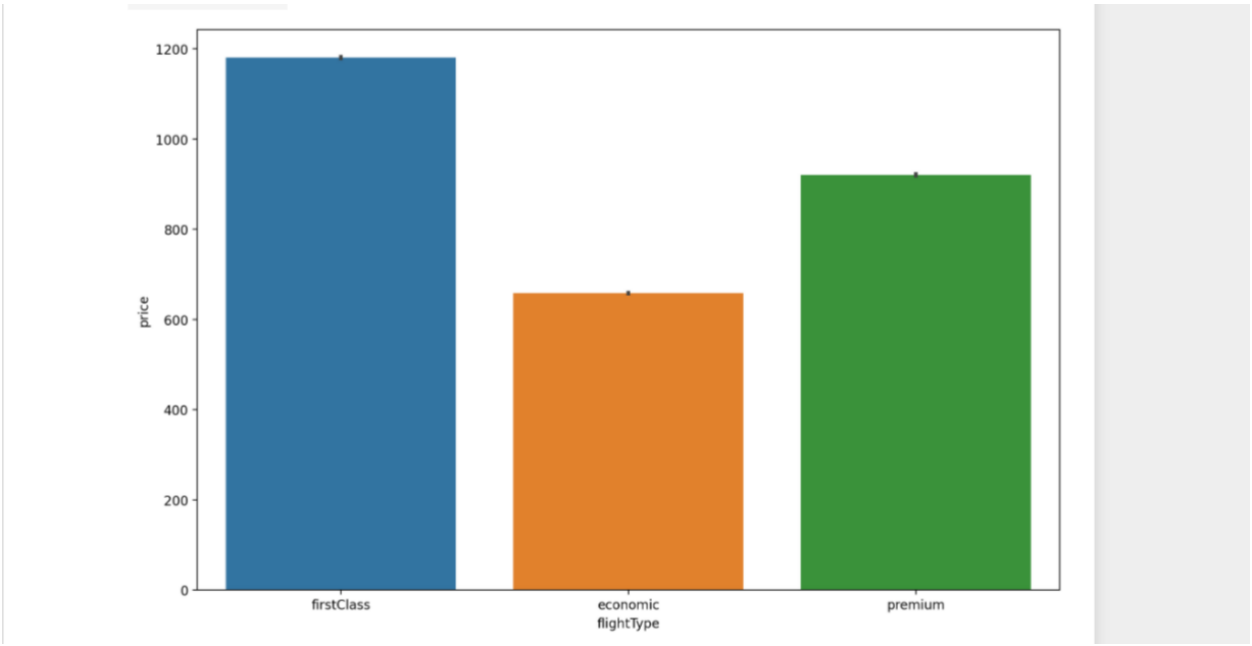
In statistics, a negatively skewed (also known as left-skewed) distribution is a type of distribution in which more values are concentrated on the right side (tail) of the distribution graph while the left tail of the distribution graph is longer.

While plotting histograms with kde for the numerical columns,I observed that-

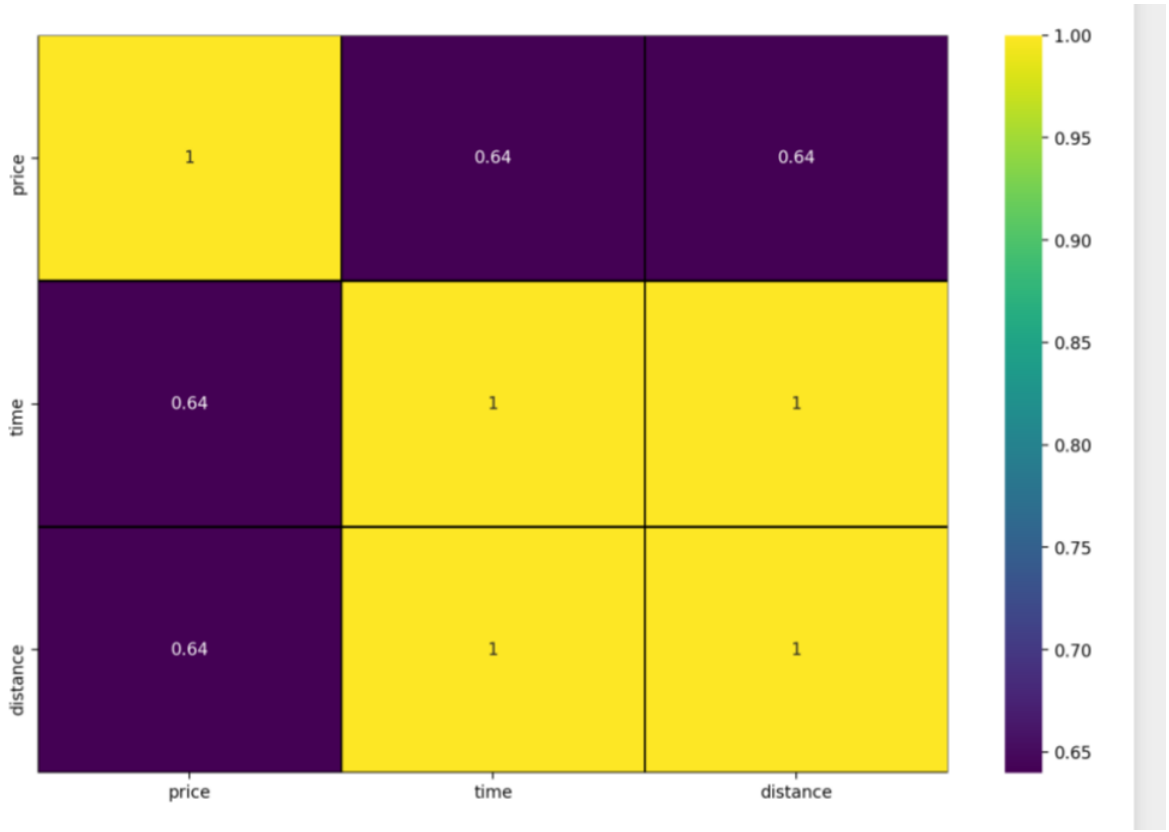
1. price -> this column distribution is right-skewed.
2. time -> this column distribution is not normal.
3. distance -> this column distribution is not normal.

2. Bar plot

   To show comparison of different data based on categorical columns.

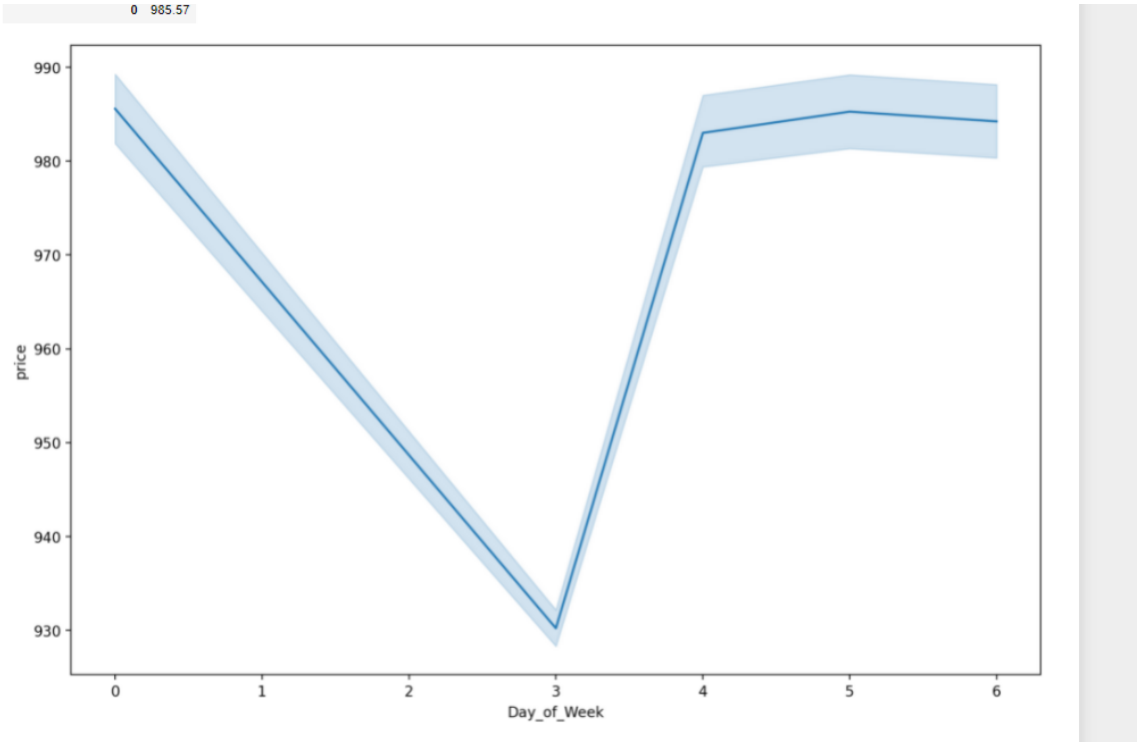


3. Heatmap
    - ☐ To show correlation between the variables and target label of the data.



- ☐ To show the chi square test results.

4. Line Plot -To show comparison of different data based on categorical columns year by year.



- Interpretation of the Results

  Give a summary of what results were interpreted from the visualizations, preprocessing and modelling.

  I found that while calculating skewness of the dataset,***all*** of the columns /variables were **approximately normally distributed**. When comparing all the models over r2 score,mean absolute error,mean squared error and root mean squared error---light gbm regression and random forest regressor were consistent. After doing a grid search cv, I found that a random forest regressor performed the best among all models. I created my final model for flight price prediction.I saved that model with the help of the **joblib** library.

CONCLUSION

- Key Findings and Conclusions of the Study

  Describe the key findings, inferences, observations from the whole problem.

  I found that the whole data was approximately normal and there were no missing values in the dataset.

- Learning Outcomes of the Study in respect of Data Science

  List down your learnings obtained about the power of visualization, data cleaning and various algorithms used. You can describe which algorithm works best in which situation and what challenges you faced while working on this project and how you overcame that.

  My answer-

  - I learned how to organize my data for the machine learning process.
  - I applied one of my favourite methods for data analysis which I often see in Excel--**pivot tables.**
  - I found out that skewness can be a great technique to know a lot about the data like--
    - ☐ whether it is normal or not,
    - ☐ how it affects outliers.
    - ☐ How it can be used to handle missing values.

- I found the seaborn heatmap plotting correlation to be better than scatterplots at finding out the relationships between the variables of the dataset.