FLIP ROBO

Project Report

On Flight Price Prediction

Submitted By: Shashanka

ACKNOWLEDGEMENT

Thefollowingresearchpapershelpedmeunderstandtheflightprices,thevariousfactorsaffectingthe pricingofaparticularflight,fluctuationsinflightpricesandtheircauses&finally,helpedmeinmymo delbuilding&predictions:

1. "<u>PredictingFlightPricesinIndia</u>"- AchyutJoshi,HimanshuSikaria, TarunDevireddy

Optimal timing for airline ticket purchasing from the consumer's perspective ischallengingprincipallybecausebuyershaveinsufficientinformationforreasoningaboutfuture pricemovements.Inthisprojectwemajorlytargetedtouncoverunderlyingtrendsofflightpricesinl ndiausinghistoricaldataandalsotosuggestthebesttimetobuyaflight ticket.

2."<u>FlightPricePrediction:ACaseStudy</u>"- PrithvirajBiswas,Rohan Chakraborty,TathagataMallik,RohanChakraborty,SkImranUdd in,ShreyaSaha,PallabiDas,SourishMitra

Alotoffactorsthataffecttheoverallpriceofairlinetickets,includingtheairline,thedateoftravel,sour ce,destination,route,duration,andsoon.Eachproviderseemstohaveitsownuniquesetregulation sandmethodsfordeterminingpricing.RecentbreakthroughsinArtificial Intelligence (AI) and Machine Learning (ML) allow for the inference of suchprinciplesaswellasthemodelingofpricevolatility.Thisarticleisastudyconductedonpredicti ngflightprices.Utilizingtwodatasetsfortestingandtraining,thisstudyanalysesvariousmachin elearningmethodsforpredictingflightprices.

INTRODUCTION

BusinessProblemFraming

TheAirlineindustryisconsideredasoneofthemostsophisticatedindustriesinusingcomplexpricingstrategies.Now-a-daysflightpricesarequiteunpredictable.Theticketpriceschangefrequently.Customersareseekingtogetthelowestpricefortheirticket, while airline companies are trying to keep their overall revenue as high aspossible.

Usingtechnologyitisactuallypossibletoreducetheuncertaintyofflightprices.Soherewewillbepredictingtheflightpricesusingefficientmachinelearningtechniques.

Whenbookingaflight,travelersneedtobeconfidentthatthey'regettingagooddeal.TheFlightPriceAnalysisAPIusesanArtificialIntelligencealgorithmtrainedonAmadeushistoricalflightbokingdatatoshowhowcurrentflightpricescomparetohistoricalfares.Moreprecisely,itshowshowacurrentflightpricesitsonadistributionofhistoricalairfareprices. Asretrievingpricemetricsthroughaggregationtechniquesandbusinessintelligencetoolsalonecouldleadtoincorrectconclusions–forexample,incaseswherethereareinsufficientdatapointstocomputespecificpricestatistics–weusedmachinelearningtoforecastprices.Thisprovidesanefficientwaytointerpolatemissingdataandpredictcoherentflightprices.

ConceptualBackgroundoftheDomainProblem

Flightpricesareunpredictable.It’smorethanlikelythatwespenthoursontheinternetresearchingflightdeals,tryingtofigureoutanairfarepricingsystemthatseemscompletelyrandomeveryday.Flightpriceappearstofluctuatewithoutreasonandlongerflightsaren’talwaysmoreexpensivethanshorterones.

ThequestionishowtoknowtheproperFlightprice,forthatIhavebuiltaMachinelearningmodelwhichcanpredicttheFlightprice.UsingvariousfeatureslikeAirline,Source, Destination,Arrivaltime,Departuretime,Stops,TravelingdateandthePriceforthesametrip.

Nowadays,thenumberofpeopleusingflightshasincreasedsignificantly.Itisdifficultforairlinestomaintainpricessincepriceschangedynamicallyduetodifferent conditions.That’swhywewilltrytousemachinelearningtosolvethisproblem.Thiscanhelpairlinesbypredictingwhatpricestheycanmaintain.Itcanalsohelpcustomerstopredictfutureflightpricesandplantheirjourneyaccordingly.

Sousingallthispreviouslyknowninformationandanalyzingthedata,Ihaveachievedagood model that has 81% accuracy. So let’s understand what all the steps we did to reachthisgoodaccuracy.

ReviewofLiterature

Severalstudiesandrelatedworkshavebeendonepreviouslytopredictflightpricesaroundtheworl dusingdifferentmethodologiesandapproaches,withvaryingresultsofaccuracyfrom50% to90%.

GiniandGroves[2]tookthePartialLeastSquareRegression(PLSR)fordevelopingamodelofpredic tingthebestpurchasetimeforflighttickets.Thedatawascollectedfrommajortraveljourney booking websites from 22February 2011to23June 2011. Additionaldatawerealsocollectedandareusedtocheckthecomparisonsoftheperformancesofthe finalmodel.

Janssen [3] built up an expectation model utilizing the Linear Quantile BlendedRegressionstrategyforSanFranciscotoNewYorkcoursewithexistingeverydayairfares given by www.infare.com. The model utilized two highlights including the number ofdays left until the takeoff date and whether the flight date is at the end of the week orweekday. The model predicts airfare well for the days that are a long way from thetakeoff date, anyway for a considerable length of time close to the takeoff date, theexpectationisn'tcompelling.

AstudybyDominguez-Menchero[16]recommendstheidealbuyingtimedependentonnonparametricisotonicrelapsem ethodforaparticularcourse,carriers,andtimeframe.Themodelgivesthemostextremenumberofd aysbeforebuyingaflightticket.twosortsof the variable are considered for the expectation. One is the passage and date ofprocurement.

Wohlfarth [15] proposed a ticket buying time enhancement model dependent on anextraordinarypre-preparingstepknownasmackedpointprocessorsandinformationminingsystems(arrangement and bunching) and measurableinvestigation strategy.Thissystemisproposedtochangeoverheterogeneousvaluearrangementinformationin toaddedvaluearrangementdirectionthatcanbebolsteredtounsupervisedgroupingcalculation.

Tziridis et al. [10] applied eight machine learning models, which included ANNs, RF,SVM,andLR,topredictticketspricesandcomparedtheirperformance.Thebestregressionmo delachievedanaccuracyof88%.

Renetal.[17]proposedusingLR,NaiveBayes,Softmaxregression,andSVMstobuildaprediction model and classify the ticket price into fivebins (60% to 80%, 80% to 100%,100%to120%, andetc.)tocomparethe relativevalues withtheoverallaverageprice.

MotivationfortheProblemUndertaken

In this project, I have to build a model that calculates the price of flight tickets, withavailableindependentvariables.Thismodelwillthenbeusedbythetravelers&tourists

tounderstandhowexactlythepricesvarywiththevariables&helpthemchoosetherightflightwiththe right price.
Thisismy4thinternshipprojectandthethirdoneinwhichIhavetobuildamachinelearning model. Also, we had to scrape the data for the model using various webscrapingtechniques.

TheFlightPricePredictionprojecthelpstravelerstofindtherightflightpricebasedon their needs and also it gives various options and flexibility for traveling.
Differentfeatures(airline,source,destination,departureandarrivaltimings,Journeydateetc.)helpstounderstandtheflightpricevariations.Usingitairlinesalsogetbenefitsandrequiredpassengers.Alsotheywillgetbenefit inscheduling also.

ANALYTICALPROBLEMFRAMING

Mathematical/AnalyticalModelingoftheProblem

Asafirststep,Ihavescrappedtherequireddatafrommakemytrip.comwebsite.Ihavefetcheddatafordifferentsourcesanddestinationsandsavedittocsvformat.
In this particular problem I have Price as my target column and it was a continuouscolumn.Soclearlyitisaregressionproblemandlhavetouseallregressionalgorithms whilebuilding themodel.

- Therewerenonullvaluesinthedataset.
- Sincewehavescrappedthedatafrommakemytrip.comwebsitetherawdatawasnotintherightformat,soIhadtousefeatureengineeringtoextracttherequiredfeature format.
- TogetbetterinsightonthefeaturesIhaveusedplottinglikedistributionplot,barplot,strip plotandcountplot.WiththeseplottingIwasabletounderstandtherelationbetweenthefeaturesinabettermanner.
- Ididnotfindanyskewnessoroutliersinthedataset.
- Ihaveusedalltheregressionalgorithmswhilebuildingmodels,thentunedthebestmodelandsavedthebestmodel.
- AtlastIhavepredictedthePriceusingthesavedmodel.

DataSources&theirformats

The data was collected from makemytrip.com website in csv format. The data wasscrapedusingselenium.Afterscrappingrequiredfeaturesthedatasetissavedasacsvfile.

Also,mydatasethad3631rowsand9columnsincludingtarget.InthisparticulardatasetI have an object type of data which has been changed as per our analysis about thedataset.Theinformationaboutfeaturesisasfollows:

★Airline:Thenameoftheairline.

★Journeydate:Thedateofthejourney★From:Thesourcefromwhichtheservicebegins.

★To:Thedestinationwheretheserviceends.

★Route:Theroutetakenbytheflighttoreachthedestination.

★DTime:Thetimewhenthejourneystartsfromthesource.

★ATime:Timeofarrivalatthedestination.

★Stops:Totalnumberofstopsbetweenthesourceanddestination.

★Price:Thepriceoftheticket

Datapreprocessingdone

- AsafirststepIhavescrappedtherequireddatausingseleniumfrommakemy trip.comwebsite.Andsavedthedataframeincsvformat.
- AndIhaveimportedrequiredlibrariesandIhaveimportedthedatasetwhichwasincsvformat.
- ThenIdidallthestatisticalanalysislikecheckingshape,nunique,valuecounts,infoetc.
- IhavealsodroppedtheUnnamed:o&Unnamed:o.1columnasIfounditwastheindexcolumnofcsvfile.
- NextasapartoffeatureengineeringIconvertedthedatatypesofdatetimecolumns andPricecolumns.
- AfterthatIsawmanysimilarvalueswhichcouldbegroupedtogetherunderacommonnamesoIreplacedallthosevaluesunderasingleheading.
- WhilecheckingfornullvaluesIfoundnonullvaluesinthedataset.
- Ihaveextractedusefulinformationfromtherawdataset.Thinkingthatthisdatawillhelp usmorethanrawdata.
- Then,Icheckthestatisticaldescriptionofourdataset.
- Lastly,Icheckedforemptyvaluespresentinourtargetcolumn,andfoundnoemptyvalues.

DataInputs-Logic-OutputRelationships

Thedatasetconsistsofatargetandotherfeatures.Thefeaturesareindependentandthetargetisdependent,asthevaluesofourindependentvariableschange,sodoesourtargetvariable change. ToanalyzetherelationbetweenfeaturesandtargetIhavedoneEDAwhereIanalyzedtherelationusingmanyplotslikebarplot,countplot,pairplot,stripplot,distplotetc. Ihavecheckedthecorrelationbetweenthetargetandfeaturesusingheatmapandbarplot, where I got the positive and negative correlation between the label and features.SinceIhadnumericalcolumnsIhaveplotteddistplotstoseethedistributionofskewnessin each column data.I have used a bar plot for each pair of categorical features thatshowstherelationbetweentargetandindependentfeatures.Ihaveusedstripplotstoseetherelationbetweennumericalcolumnsandtargetcolumns.Icannoticethereisagoodrelationshipbetweenmaximumcolumnsandtarget.

ImportantfeaturesthataffectPricepositivelyandnegativelyareasfollows:

Featureshavinghigh Positivecorrelationwithtarget: From,To.

Featureshavinghigh Negativecorrelationwithtarget: Stops,Airline.

HardwareandSoftwareRequirements&Toolsused

Tobuildthemachinelearningprojectsitisimportanttohavethefollowinghardwareandsoftwarerequirementsandtools.

Hardwarerequired:

- Processor:corei5orabove
- RAM:8GBorabove
- ROM/SSD:250GBorabove

Softwarerequired:

- Anaconda-languageusedPython3

LibrariesUsed:

- importnumpyasnp
- importpandasaspd
- importseabornassns
- importmatplotlib.pyplotasplt
- fromsklearn.preprocessingimportLabelEncoder
- fromsklearn.preprocessingimportStandardScaler
- fromstatsmodels.stats.outliers_influenceimportvariance_inflation_factor
- fromsklearn.linear_model.LinearRegression
- fromsklearn.treeimportDecisionTreeRegressor
- fromsklearn.neighbors.KNeighborsRegressor
- fromsklearn.svm.SVR
- fromsklearn.ensembleimportRandomForestRegressor
- fromxgboostimportXGBRegressor
- fromsklearn.ensembleimportGradientBoostingRegressor
- fromsklearn.ensembleimportExtraTreesRegressor
- fromsklearn.model_selectionimportcross_val_score
- fromsklearn.model_selectionimportGridSearchCV

DATAANALYSIS&VISUALIZATION

Identificationofpossibleproblem-solvingapproaches(methods)

- Sincethedatacollectedwasnotintherightformatwehadtocleanitandbringittotheproperformatforouranalysis.
- Therewerenooutliersandskewnessinthedataset.
- Wehavedroppedalltheunnecessarycolumnsinthedatasetaccordingtoourunderstanding.
- UseofPearson'scorrelationcoefficienttocheckthecorrelationbetweendependent andindependentfeatures.
- Next,weencodedthevariouscategoricalcolumnsusingLabelEncoder.
- AlsoIhaveusedStandardizationtoscalethedata.
- AfterscalingwehavetocheckmulticollinearityusingVIF.
- ThenfollowedbymodelbuildingwithallRegressionalgorithms.

TestingofIdentifiedApproaches(Algorithms)

Since Price was my target and it was a continuous column with improper format whichhad to be changed to continuous float data type column, this particular problem was aRegressionproblem.AndIhaveusedallRegressionalgorithmstobuildmymodel.
By looking into the r2 score and error values I found RandomForestRegressor as a bestmodelwithhighestr2_scoreandleasterrorvalues.
BelowisthelistofRegressionalgorithmsIhaveusedinmyproject:

- LinearRegression
- SVR
- KNearestNeighborsRegressor
- RandomForestRegressor
- XGBRegressor
- ExtraTreesRegressor
- GradientBoostingRegressor
- DecisionTreeRegressor

KeyMetricsforsuccessinsolvingproblemunderconsideration

Ihaveusedthefollowingmetricsforevaluation:

- Ihaveusedmeanabsoluteerrorwhichgivesamagnitudeofdifferencebetweenthepredictionofanobservationandthetruevalueofthatobservation.

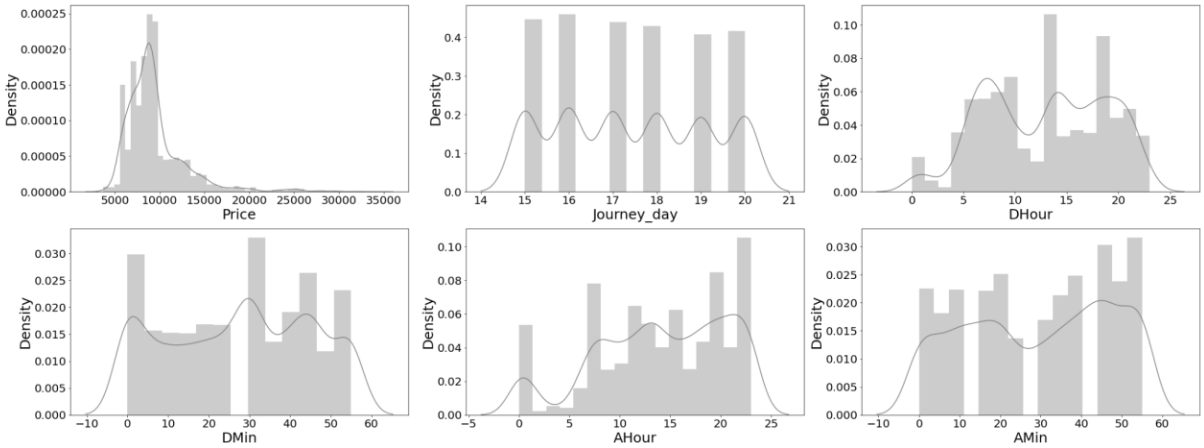
- Ihaveusedrootmeansquaredeviationasoneofthemostcommonlyusedmeasure sforevaluatingthequalityofpredictions.
- IhaveusedtheR2scorewhichtellsushowaccurateourmodelis.

Visualizations

I have used count plots for the categorical features & bar plots to see the relation ofcategorical variables with the target and I have used 2 types of plots for numericalcolumnsoneisdistplotforunivariateandstripplotforbivariateanalysis.

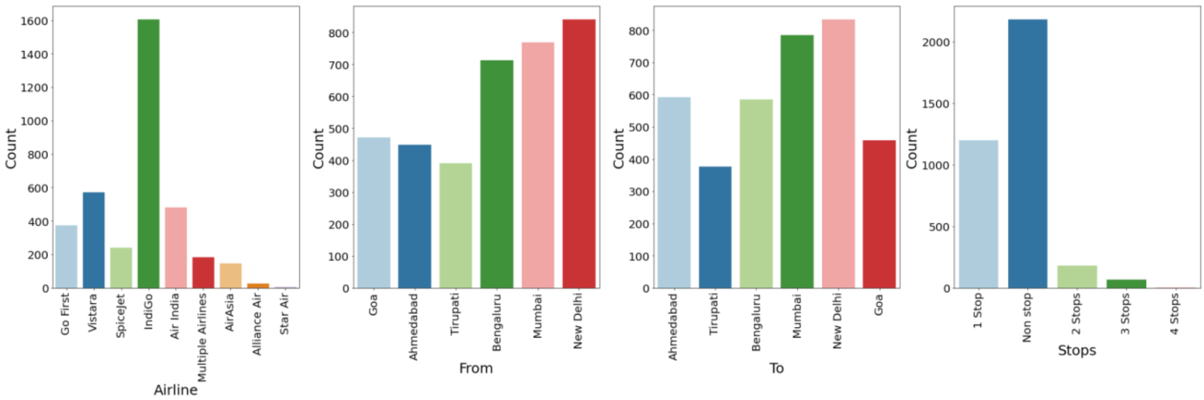
<u>UNIVARIATEANALYSIS:</u>

Numericalcolumns:



<u>OBSERVATION</u>:Thereisnoskewnessinanyofthenumericalcolumns.

Categoricalcolumns:



<u>OBSERVATIONS</u>:

<u>Airline</u>:IndigohasmaximumcountwhichmeansmostofthepassengerspreferredIndigofortheirtr aveling.

<u>From</u>:NewDelhihasmaximumcountforsourcewhichmeansmaximumpassengersarechoosingNe wDelhiastheirsource.

<u>To</u>:NewDelhihasmaximumcountforDestinationwhichmeansmaximumpassengersarechoosi ngNewDelhiastheirDestination.
<u>Stops</u>:MostoftheflightshaveNostopsinbetweenandaredirectflights.Secondly,mostflightshave1 stopinbetween.

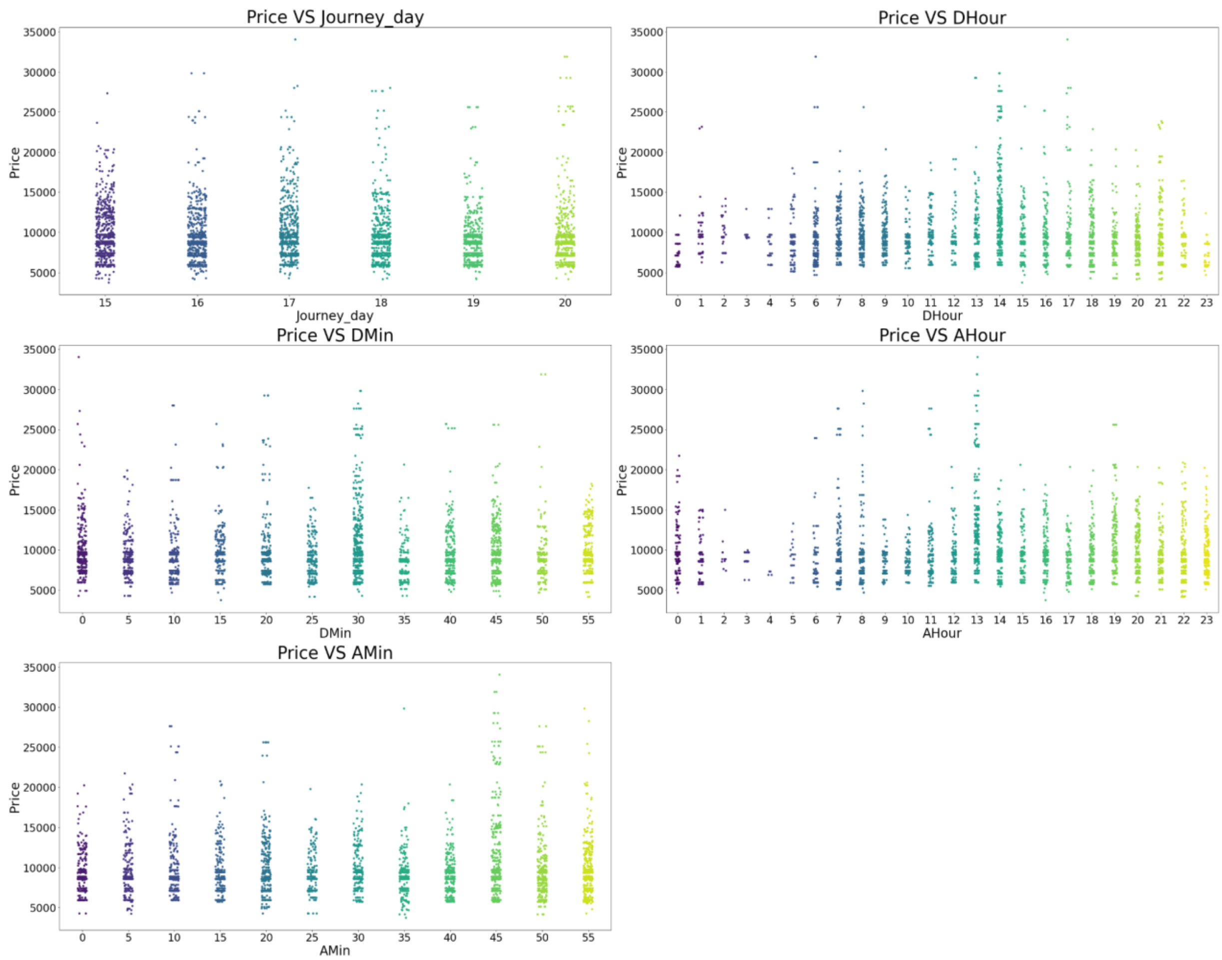
<u>BIVARIATEANALYSIS</u>:

Categoricalcolumns:

<u>OBSERVATIONS</u>:

<u>PricevsAirline</u>-ForMultipleAirlinesthePriceishighcomparedtootherAirlines.<u>PricevsFrom</u>-TakingTirupatias SourcecoststhehighestPriceComparedtootherSourcepoints.
<u>PricevsTo</u>-TakingTirupatiasDestinationcostshighestPriceComparedtootherDestinationpoints.
<u>PricevsStops</u>-4stopsorhigherarecostlier,comparedtonostops&1stop.

Numericalcolumns:



<u>OBSERVATIONS</u>:

<u>PricevsJourney_day</u>-Inallthedatesthepriceisalmostthesame.

<u>PricevsDHour</u>-
At2PMdeparturetimeofeverydaytheflightPricesarehighsoitlooksgoodtobookflightsratherthan thisdeparturetime.

<u>PricevsDMin</u>-AndDepartureminutehaslessrelationwithtargetPrice.

<u>PricevsAHour</u>-
At6AM,7AM,8AM,11AM&1PMArrivaltimeofeverydaytheflightPricesarehighsoitlooksgood tobookflightsratherthanthisarrivaltime.<u>PricevsAMin</u>-
ArrivalminutehaslessrelationwithtargetPrice.

MODELBUILDING&PREDICTION

RunandEvaluateselectedmodels

<u>RegressionModels:</u>

LINEARREGRESSION:

```
In [71]: from sklearn.linear_model import LinearRegression

lr=LinearRegression()
lr.fit(x_train,y_train)
lr.score(x_train,y_train)

pred_lr=lr.predict(x_test)
print('R2_Score: ',r2_score(y_test,pred_lr))
print('Mean absolute error: ',mean_absolute_error(y_test,pred_lr))
print('Mean squared error: ',mean_squared_error(y_test,pred_lr))
print('Root Mean squared error: ',np.sqrt(mean_squared_error(y_test,pred_lr)))

R2_Score:  0.1897812736568698
Mean absolute error:  2101.264726904694
Mean squared error:  10953837.20037541
Root Mean squared error:  3309.65816971714
```

TheLinearRegressionmodelgaveusanR2Scoreof18.97%.DECISIONTREER

EGRESSOR:

```
In [72]: from sklearn.tree import DecisionTreeRegressor

dtr = DecisionTreeRegressor()
dtr.fit(x_train,y_train)
dtr.score(x_train,y_train)

pred_dtr=dtr.predict(x_test)
print('R2_Score: ',r2_score(y_test,pred_dtr))
print('Mean absolute error: ',mean_absolute_error(y_test,pred_dtr))
print('Mean squared error: ',mean_squared_error(y_test,pred_dtr))
print('Root Mean squared error: ',np.sqrt(mean_squared_error(y_test,pred_dtr)))

R2_Score:  0.6859877947695018
Mean absolute error:  987.4315886134068
Mean squared error:  4245321.001836548
Root Mean squared error:  2060.417676549235
```

TheDecisionTreeRegressorModelgaveusaR2Scoreof68.59%.KNEARESTN

EIGHBORSREGRESSOR:

```
In [73]: from sklearn import neighbors

knn = neighbors.KNeighborsRegressor()
knn.fit(x_train,y_train)
knn.score(x_train,y_train)

pred_knn=knn.predict(x_test)
print('R2_Score: ',r2_score(y_test,pred_knn))
print('Mean absolute error: ',mean_absolute_error(y_test,pred_knn))
print('Mean squared error: ',mean_squared_error(y_test,pred_knn))
print('Root Mean squared error: ',np.sqrt(mean_squared_error(y_test,pred_knn)))

R2_Score:  0.6813181031806808
Mean absolute error:  1356.6927456382002
Mean squared error:  4308453.387915519
Root Mean squared error:  2075.68142736681
```

TheKNearestNeighborsRegressionModelgaveusaR2Scoreof68.13%.SUPPORTVE CTORREGRESSOR(SVR):

```
In [74]: from sklearn.svm import SVR

svr=SVR()
svr.fit(x_train,y_train)
svr.score(x_train,y_train)

pred_svr=svr.predict(x_test)
print('R2_Score: ',r2_score(y_test,pred_svr))
print('Mean absolute error: ',mean_absolute_error(y_test,pred_svr))
print('Mean squared error: ',mean_squared_error(y_test,pred_svr))
print('Root Mean squared error: ',np.sqrt(mean_squared_error(y_test,pred_svr)))

R2_Score:  -0.06085742829601459
Mean absolute error:  2240.540971724598
Mean squared error:  14342373.466004254
Root Mean squared error:  3787.132617958375
```

TheSVRModelgaveusaR2Scoreof-

6.08%.RANDOMFORESTREGRESSOR:

```
In [75]: from sklearn.ensemble import RandomForestRegressor

rfr = RandomForestRegressor()
rfr.fit(x_train,y_train)
rfr.score(x_train,y_train)

pred_rfr=rfr.predict(x_test)
print('R2_Score: ',r2_score(y_test,pred_rfr))
print('Mean absolute error: ',mean_absolute_error(y_test,pred_rfr))
print('Mean squared error: ',mean_squared_error(y_test,pred_rfr))
print('Root Mean squared error: ',np.sqrt(mean_squared_error(y_test,pred_rfr)))

R2_Score:  0.8163277124591364
Mean absolute error:  858.1481154619792
Mean squared error:  2483176.789832809
Root Mean squared error:  1575.809883784465
```

TheRandomForestRegressionModelgaveusaR2Scoreof81.63%.GRADIENTBOOS

TINGREGRESSOR:

```
In [76]: from sklearn.ensemble import GradientBoostingRegressor

gbr=GradientBoostingRegressor()
gbr.fit(x_train,y_train)
gbr.score(x_train,y_train)

pred_gbr=gbr.predict(x_test)
print('R2_Score: ',r2_score(y_test,pred_gbr))
print('Mean absolute error: ',mean_absolute_error(y_test,pred_gbr))
print('Mean squared error: ',mean_squared_error(y_test,pred_gbr))
print('Root Mean squared error: ',np.sqrt(mean_squared_error(y_test,pred_gbr)))

R2_Score:  0.7485119744115258
Mean absolute error:  1154.6295506731049
Mean squared error:  3400018.7857585284
Root Mean squared error:  1843.9139854555388
```

TheGradientBoostingRegressorModelgaveusaR2Scoreof74.85%.

EXTRATREESREGRESSOR:

```
In [77]: from sklearn.ensemble import ExtraTreesRegressor

         etr=ExtraTreesRegressor()
         etr.fit(x_train,y_train)
         etr.score(x_train,y_train)

         pred_etr=etr.predict(x_test)
         print('R2_Score: ',r2_score(y_test,pred_etr))
         print('Mean absolute error: ',mean_absolute_error(y_test,pred_etr))
         print('Mean squared error: ',mean_squared_error(y_test,pred_etr))
         print('Root Mean squared error: ',np.sqrt(mean_squared_error(y_test,pred_etr)))

         R2_Score:  0.8128817927484211
         Mean absolute error:  807.9719237832874
         Mean squared error:  2529764.2634241735
         Root Mean squared error:  1590.5232671747287
```

The Extra Trees Regressor Model gave us a R2 Score of 81.28

%.XGBOOSTREGRESSOR:

```
In [78]: from xgboost import XGBRegressor

         xgb=XGBRegressor()
         xgb.fit(x_train,y_train)
         xgb.score(x_train,y_train)

         pred_xgb=xgb.predict(x_test)
         print('R2_Score: ',r2_score(y_test,pred_xgb))
         print('Mean absolute error: ',mean_absolute_error(y_test,pred_xgb))
         print('Mean squared error: ',mean_squared_error(y_test,pred_xgb))
         print('Root Mean squared error: ',np.sqrt(mean_squared_error(y_test,pred_xgb)))

         R2_Score:  0.8135892888736583
         Mean absolute error:  906.873558023416
         Mean squared error:  2520199.1952224993
         Root Mean squared error:  1587.5135259967076
```

TheXGBoostRegressorModelgaveusaR2Scoreof81.35%.

From the above regression models, the highest R2 score belongs to the RandomForest Regressor Model, followed closely by XGBoost Regressor Model & Extra TreesRegressorModel.Next,theGradientBoostingRegressorModel.
After that, the KNearest Neighbors Regression Model & Decision Tree Regressor Model.Lastly,theLinearRegressionModel.
Finally,thelowestscorebelongstotheSVRModel.

<u>HyperParameterTuning:</u>

SincetheR2Scoreis theHighest andtheRMSEscoreisthelowestinRandom ForestRegressorModel,weshallconsideritforhyperparametertuning.

WeshalluseGridSearchCVforHyperParameterTuning.

```
In [79]: from sklearn.model_selection import GridSearchCV
```

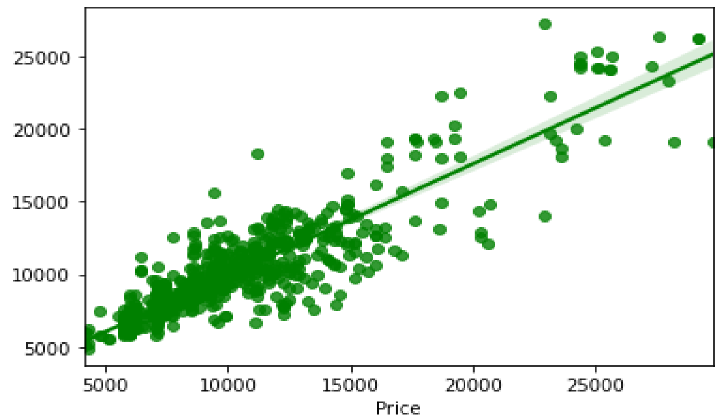
```
In [80]: parameters={
             'criterion': ['squared_error','absolute_error'],
             'max_depth': [10,20,30],
             'max_features': ['sqrt','log2'],
             'n_estimators': [100,200,300]}
         grid_rfr = GridSearchCV(rfr, param_grid = parameters, cv = 5)
```

```
In [81]: grid_rfr.fit(x_train, y_train)

Out[81]: GridSearchCV(cv=5, estimator=RandomForestRegressor(),
                      param_grid={'criterion': ['squared_error', 'absolute_error'],
                                  'max_depth': [10, 20, 30],
                                  'max_features': ['sqrt', 'log2'],
                                  'n_estimators': [100, 200, 300]})
```

```
In [82]: grid_rfr.best_params_

Out[82]: {'criterion': 'squared_error',
          'max_depth': 20,
          'max_features': 'log2',
          'n_estimators': 300}
```

```
In [112]: flight_model = RandomForestRegressor(criterion='squared_error',max_features='log2',max_depth=20,n_estimators=300)
          flight_model.fit(x_train,y_train)

          pred = flight_model.predict(x_test)
          print('R2_Score: ',r2_score(y_test,pred))
          print('Mean absolute error: ',mean_absolute_error(y_test,pred))
          print('Mean squared error: ',mean_squared_error(y_test,pred))
          print('Root Mean squared error: ',np.sqrt(mean_squared_error(y_test,pred)))

          R2_Score:  0.8149575169163876
          Mean absolute error:  896.3206010910508
          Mean squared error:  2501701.2924392736
          Root Mean squared error:  1581.6767344938958
```

AfterHyperParameterTuning,wehavegotaR2scoreof81.49%.



Savingthefinalmodelandpredictingthesavedmodel

Nowweshallsavethebestmodel.

```
In [114]: import joblib
          joblib.dump(flight_model,"Flight_Price_Prediction_Model.pkl")

Out[114]: ['Flight_Price_Prediction_Model.pkl']
```

```
In [115]: # Loading the saved model
          flight_price_model=joblib.load("Flight_Price_Prediction_Model.pkl")

          # Prediction
          prediction = flight_price_model.predict(x_test)
          prediction

Out[115]: array([ 6001.35333333,  8701.97666667,  7086.4       , ...,
                 14868.03      ,  8612.97666667,  9047.88333333])
```

Puttingthepredicted&actualvaluesinadataframe.

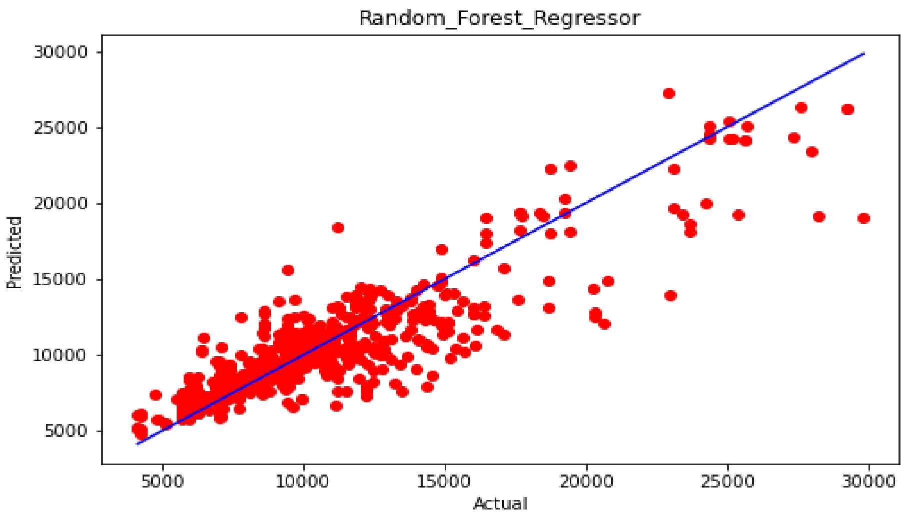
```
In [116]: pd.DataFrame([flight_price_model.predict(x_test)[:],y_test[:]],index=["Predicted","Actual"])
```

Out[116]:

	0	1	2	3	4	5	6	7	8	9	...	1079	1080
Predicted	6001.353333	8701.976667	7086.4	9271.646667	13887.37	9417.62	5818.873333	13789.773333	9757.893333	8464.743067	...	8137.72	19092.042667
Actual	5943.000000	8567.000000	7056.0	9314.000000	11543.00	8839.00	5943.000000	13830.000000	9679.000000	9597.000000	...	8346.00	16480.000000

2 rows × 1089 columns

Graphofpredictedandactualvalues.



InterpretationoftheResults

Thedatasetwasscrappedfrommakemytrip.comwebsite.Thedatasetwasverychallengingtohandle;ithad9featureswith3631samples.

- Firstly,thedatasetshad2completerowsasnanvalues,soIhavedroppedthoserows.
- Andtherewasahugenumberofunnecessaryentriesinallthefeaturessolhaveusedfeatur eextractiontogettherequiredformatofvariables.
- Andproperplottingforthepropertypeoffeatureswillhelpustogetbetterinsighton the data. I found both numerical columns and categorical columns in thedatasetsoIhavechosenstripplotandbarplottoseetherelationbetweentargetandfeat ures.
- Ididnotfindanyoutliersorskewnessinthedataset.
- IthenencodedthedatasetwithLabelEncoder.
- Thenscalingthedatasethasagoodimpactlikeitwillhelpthemodelnottogetbiased.Since wedidnothaveoutliersandskewnessinthedatasetsowehavetochooseStandardizat ion.
- Weusedmultiplealgorithmswhilebuildingamodelusingthedatasettogetthebestmod el outofit.
- Andwehavetousemultiplemetricslikemse,mae,rmseandr2_scorewhichwillhelpusto decidethebestmodel.
- IfoundRandomForestRegressortobethebestmodelwith81.49%r2_score.
- AlsoIhaveimprovedtheaccuracyofthebestmodelbyrunninghyperparametertuning.
- AtlastIhavepredictedtheusedflightpriceusingthesavedmodel.Iwasabletogetthepred ictionsneartoactualvalues.

CONCLUSION

KeyFindingsandConclusionsoftheStudy

Inthisproject,wehaveusedmachinelearningalgorithmstopredicttheflightprices.Wehav ementionedthestepbystepproceduretoanalyzethedatasetandfindthecorrelationbetweenthefe atures.Thuswecanselectthefeatureswhicharecorrelatedtoeachotherandareindependentin nature.

Thesefeaturesetswerethengivenasaninputto8algorithmsandahyperparametertuningw asdonetothebestmodelandtheaccuracyhasbeenimproved.

Thus, we calculated the performance of each model using different performancemetrics and compared them based on those metrics. Then we have also saved the bestmodelandpredictedtheflightprice.

Itwasgoodthatthepredictedandactualvalueswerealmostthesame.

LearningOutcomesoftheStudyinrespectofDataScience

Ifoundthatthedatasetwasquiteinterestingtohandleasitcontainsalltypesofdataandthedatawasscrapedfrommakemytrip.comwebsiteusingselenium.

New analytical techniques of machine learning can be used in flight priceresearch. The power of visualization has helped us in understanding the data bygraphical representation. It has made me understand what data is trying to say. Datacleaningisoneofthemostimportantstepstoremoveunrealisticvaluesandnullvalues.

Thisstudyisanexploratoryattempttouse8machinelearningalgorithmsinestimatingflightpriceprediction,andthencomparetheirresults.

We hope this study has moved a small step ahead in providing somemethodologicalandempiricalcontributionstocreditingonlineplatforms,andpresentinganalternativeapproachtothevaluationofflightprice.

LimitationsofthisworkandScopeforFutureWork

LIMITATIONS:

- Firstdrawbackisscrapingthedataasitisafluctuatingprocess.
- Followedby rawdatawhich isnot informat toanalyze.
- Also, we have tried best to deal with improper format data and null values. So itlooksquitegoodthatwehaveachievedanaccuracyof81.49%evenafterdealingwith allthesedrawbacks.
- Also,thisstudywillnotcoverallRegressionalgorithms;instead,itisfocusedonthechosenalgorithm.

FUTUREWORK:

- Futuredirectionofresearchmayconsiderincorporatingadditionalusedflightdatafromalargereconomicalbackgroundwithmorefeatures.