International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.140

Airfare Estimator Using Random Forest Algorithm

Vikyath Kumar M S
Department of Data Science and Business Systems
SRM Institute Of Science And Technology
Kattankulathur, Chennai, India
vk8872@srmist.edu.in

Vikyath Kumar M S
Department of Data Science and Business Systems
SRM Institute Of Science And Technology
Kattankulathur, Chennai, India
vk8872@srmist.edu.in

Anand Madasamy
Department of Data Science and Business Systems
SRM Institute Of Science And Technology
Kattankulathur, Chennai, India
anandm4@srmist.edu.in

***Abstract*—Nowadays people have started to prefer Air Trans-port compared too there mode so transportation. Thereasonforthepublictooptforthismodeoftransportisitisthefastest mode of travel and it also provides more comfort, safety, organized system, staff support during the journey etc...One of the major problems faced by the public regarding air transportisthatthefareofflightticketskeepsfluctuatingsignificantlyand dynamically. The airline companies are basically one of the most subtle companies in making complex pricing schemes. They usually increase the price when the demand is high. There are multiple factors that affect the price of the flight ticket such as duration of the journey, source, arrival time, departure time and so on. Usually, using the past data, time series analysis is done manually to get the estimation of a flight ticket in the future. To make this process easier and simpler we have created a web application which uses machine earning algorithms to predict the flight fare based on the previous data which has been collected.WecollectedourdatasetfromKaggleandappliedMLalgorithmsand regression techniques to obtain the results. Python languageisbeingusedtodevelopmachinelearningmodules.Wehaveuseda random forest algorithm in our project to predict the flight farebased upon the historical data available. To optimise the model,wehaveperformedHyperparametertuningtogetthebestresultswith higher accuracy. To provide a better user experience to theuser we created a web application using Flask where users cangive the inputs and obtain the results on the screen. This servicehelpstheuserstobooktheirflightticketsatalowerprice.**

***Keywords*—MachineLearning(ML),Airfare,RandomFor-est(RF),Hyper-parameterTuning**

I. INTRODUCTION

The main aim of the airline industry is to increase theirprofit for which they sell flight tickets at a higher cost, sellmore tickets, and many other strategies. But on the other hand,the customer's goal is to buy the ticket at a lower price. Theflightfareofaparticularflightmayvaryupto7timesadayaspertheresearchers.However,differencesinpassengerdemandand available seats usually lead to customers purchasing theticket for a higher cost or might cause a revenue loss to airlineindustries.Usually,airlinecompaniesaremostlyequippedwith advanced tools, capabilities, and a team to control thepricingprocess.Butforacommonman,it'snotthateasytoestimatetheprice.Frequentlytravelingpeoplehaveanapproximateidearegardingwhentobooktheflightticket togetitatthebestprice. Butmanyinexperiencedpeopleland into the traps of discounts made by the companies andfinally end up paying more than the actual cost. Therefore, ourproposed system can help millions of people in society to savemoney by providing them with detailed information regardingthe right time to book a flight ticket. For determining the pricewe need some features as input such as the duration of thejourney,source,arrivaltime,departuretime,andsoon.

Ourproposedproblemstatementis"AirwaysFareEstimatorUsingRandomForestAlgorithm".

II. AIMSANDOBJECTIVES

A. Aims

1. Togetbetterexposureandknowledgeinthefieldofdatascienceandmachinelearning.
2. Togetthebestpriceofairfarewithgoodaccuracy.
3. Provideauser-friendlyinterfaceandthebestuserexperiencewhileusingthewebapp.
4. Tostudyadetailedanalysisofthefactorsthatinfluenceairfarecost.

B. Objectives

1. ProvideabetterUserInterfacetotheusers.
2. Togetthebestresultswiththebestpossibleaccuracy.
3. Use machine learning models to train the data-set, and getaccurateresultsalongwithbetterruntimesothattheusergetsthebestuserexperiencewhileusingthewebapplication.

III. LITERATURE SURVEY

They used a data-set that had information generated from1814 distinct flights that were performed by Aegean Airlinesforthepurposeofthestudy[1],anditwasfoundthatthestudy was successful. Using the data that was supplied, theytaughtthemachine-learningmodelhowtobehave.Inorderto illustrate that the selection of features may influence theefficacy of a machine learning model, they tested the modelusing a broad variety of attributes in addition to the collecteddata. This allowed them to demonstrate that the performanceofthemodelcouldbeaffectedbythechoiceoffeatures.

Theauthorsoftheresearchpaper[2]usedasmallerdata-settodeveloppredictionsabouttravelfaresusingmachinelearningalgorithms.Thesealgorithmswererunusing the data. The information gathered from this collectionof data contained details on each and every flight that travelsbetweenBombayandDelhi.Inorderforthemtofinishtheir study,theyemployedanumberofdifferentmachinelearningstrategies,suchasK-nearestneighbors(KNN),linear regression, and SVM-Support Vector Machine. Thesestrategieswerealleffective.

The researchers were able to materialise the model that theyhad conceived of thanks to the application of a methodologyknownasLinearQuantileBlendedRegression(LQBR),which was utilised in the study [3]. For the purpose of thisstudy, a data set was utilised that contained 126,412 individualobservationsonthecostofaticketforeachofthe2,271i

DOI: rp-9788770040723.140

ndividual flights that took place between the San FranciscoInternational Airport and the New York International Airport.As a means of determining quality, these observations werecarriedoutonadailybasisasthestandard.

Intheresearcharticlepublished[4],theauthorspresentedam odelinwhichthetwodatabases,inadditiontothemachine learning techniques and the macroeconomic data, aremerged. Based on the source and destination data, machinelearningmethodssuchasXGboostandSVM(SupportV ector Machine) are utilised in order to make a predictionregardingtheairfare.Afterperformingminoradjustm entsto the R-squared performance measurements, the suggestedframework is able to produce prediction results with a greaterlevel of accuracy. Using the XGBoost Algorithm, they wereable to attain an error rate that was far lower than average,cominginatapproximately0.92.

Usingmachinelearningalgorithmsonflightdatasetsenable stheforecastingofdynamicflightfaresanddeterminingthe most favorable ticket prices. As the data is sourced fromwebsitessellingflighttickets,theavailableinformationisre stricted.R-squaredvaluesareutilizedtoevaluatetheprecisionofthemodel.I ncorporatingsupplementarydata,suchasthepresentseatavailab ility,couldenhancetheaccuracyofthepredictions.Theprocessof predictingflightcostshasbeenexhaustivelydescribed,andprior patternshavebeenemployedtoconfirmthecredibilityofthesepr ojections.[5]

Therandomforesttechniqueisasimpleandadaptablealgorit hmthatcanimproveaccuracyandprovideflexibilityinsolvingav arietyofclassificationandregressiontasks.Decision trees, which are trained on different subsets of thedata,arepartoftherandomforestmodel.Bycombiningmultip le decision trees and reducing the negative impact ofbias and variance, the random forest method typically deliversbetterresults.[6]

The study has shown that incorporating dynamic pricinginto an airline's revenue management system can result in asignificant revenue boost in comparison to traditional revenuemanagement methods. Dynamic pricing can yield short-termrevenuegainsofupto20percent,owingtoitssuperiorflexibil ity in responding to changes in the environment. Thisflexibilityismainlyduetothefactthatdynamicpricingtechni ques do not establish a fixed booking control policy atthe outset of the booking period, as opposed to static methods.Nonetheless,therevenuebenefitofdynamicpricingma ydeclineorbebalancedoutinthelongrun,ascompetitorsalsoado ptcomparablestrategies.[7]

This research proves that it is feasible to make use of pastdatatoanticipatethecostofairfare.Torefinetheaccuracyof the forecasts, one possible strategy is to merge differentmodels and assess their efficacy for each category. The curveoflearningimpliesthatincorporatingmorecharacteristics would raise the model's precision even further. Nevertheless,due to the limitations of our existing data source, we cannotextractmoreinformationonparticularflights.Movingforward, more characteristics, like seat availability, departuretime, and holiday schedules, may be included in the model toboostitspredictingcapacity.[8]

IV. System Architecture

The data-set we used has 10,000+ observations along withthe booking details namely Airline(Company), Journey Date,Destination, Source, Arrival-Time, Departure-Time, Durationof the Journey, Total number of stops, additional information,and finally the prices which act as our target variable. FeatureEngineering is performed to convert all the above-mentionedfeaturestonumericalrepresentation.Later,tofinalize thetraining model we use VIF Multicollinearity and Sklearn -Feature importance. After completion of the above two stages,weperformmodeltrainingusinganappropriateAlgorith mthatprovides the best results according to our objectives. Finally,we deploy our model using the Flask services. Therefore, wecanrunourwebappanddeployitinaliveenvironmentfor real-time usage. Figure 1 shows the overview of systemarchitectureofairfareestimatorusingmachinelearningm odel

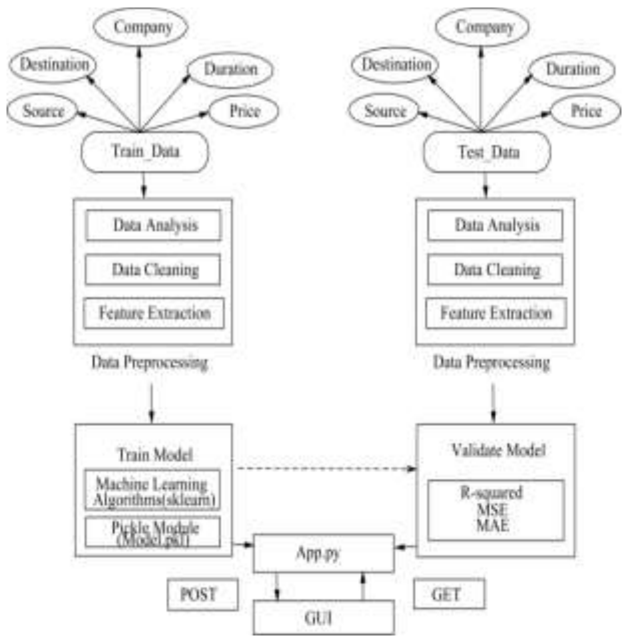

Fig.1.System Architecture

V. Proposed Methodology

The study has been designed to include seven main stages,each with a unique objective. Figure 2 represents the flowdirection of all the seven phases in the proposed methodology.Each of these processing phases is described in greaterdetailinthefollowingsections.

Phase1:ImportingLibraries:Allthenecessarypythonlibrar iesrequiredforAirwaysFareEstimatorareimportedthrough python commands. Some such libraries are Numpy,Pandas,andsoon.

Phase 2: Data Selection: Initially a proper dataset has to becollectedonwhichthedatatraininghastobeperformed.Forthis purpose,mostmachinelearningexperts,students,researchers,sc holars,anddatascientistsusetheKaggleservicewhich is a

DOI: rp-9788770040723.140

subsidiary of Google. The dataset collected shouldbe loaded by setting up the working directory for performingfurtheractions.

Phase3:EDA:Itistheprocessofunderstandingthedata.Itis used to analyze the trends and statistical summary in theformofagraphicalrepresentation.

Phase4:DataPreprocessing:Datapreprocessingplaysacritical role in the fields of data analysis and machine learningas it entails the task of refining, reformatting, and arrangingunprocessed data to render it appropriate for advanced analysis.Theprimaryobjectiveofdatapreprocessingistoguaranteethat the data is precise, coherent, and in a machine-readableformat.

Phase 5: Feature Selection: This is the process of finding outthebestfeatureamongtheavailablefeaturesthatwillleadto maintaining a good relationship with the target variable.Feature importance and VIF-Multicollinearity are the methodsthatwillbeusedintheprojecttoperformthisprocess.

Phase6:ModelTraining:Thisistheprocesswherethecollected data is used to train the model with the help of MLalgorithms.ThemodelcanpredictthepriceusingthehistoricaldataandbyperformingthedatatrainingwithanappropriateMLalgorithm,thereforeachievingthebestresults.

Phase 7: Deployment of Model: The machine learning modelcreated will be deployed as a web app using flask services.Using this the user can interact with the model for entering theinput values for which the airfare should be predicted. Finally,theresultobtainedisalsodisplayedontheuserinterfaceofthewebapp.Thiswebappensuresabetteruserexperience.

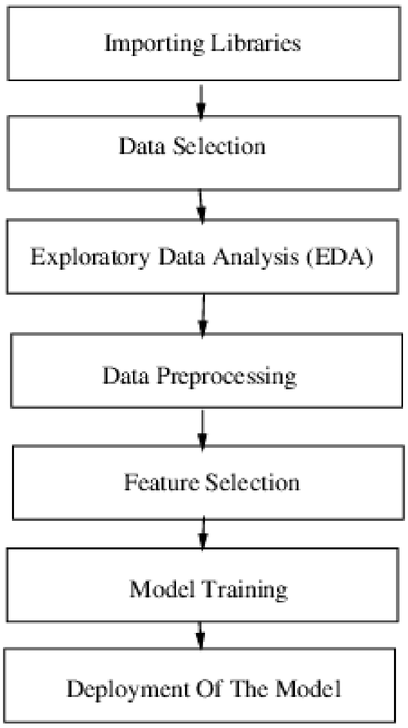

Fig.2.ProposedMethodology

VI. Model Training

Modeltrainingisacrucialprocessinmachinelearningthat involves the development of a mathematical model usingan algorithm to learn from input data. The ultimate goal ofmodeltrainingistocreateanaccurateandrobustmodelthat can provide precise predictions or classifications for newand unseen data.To begin the model training process, a largeamount of data is collected and preprocessed through severalsteps like data cleaning, normalization, and feature extractionto make it ready for training. Afterward, the model selectionprocess starts, where an appropriate algorithm is chosen basedontheproblem’snatureandthetypeofdata.Inordertoperform model training, an in-depth experimental and researchanalyseswasdoneandfinally,weoptedRandomForestalgorithmtotrainourmodel.Performancemetricsareasetof standards or markers that are utilized to assess the efficacy,productivity,andoverallperformanceofasystem,process,orproduct.Weobtainedperformancemetricsscoresbasedonourexperimentaswellasfromtheresearchanalyses.TABLEIdemonstratestheresultsobtainedbyperformingthe prediction using different models.We got results for RFAlgorithm, K - Neighbours, Decision Tree and Extra TreeModels.Basedonthescoresobtaineditwasdemonstartedthat RF Algorithm performed well with a better RMSE scoreof1773.43,R-squareobtainedwas0.84andtotaltimetakentotrainthemodelwas2.51seconds.

TABLEI: Peformace Metrics Of Ml Models

Model	R-Squared	RMSE	Time Taken
Random Forest	0.84	1773.43	2.51
K - Neigh- bours	0.84	1838.12	1.43
Decision Tree	0.84	1838.85	0.09
Extra Tree	0.80	2081.43	0.06

VII. Random Forest Algorithm

This algorithm comes under the category of supervised learning. This is also used for both regression and classification problems.The random forest algorithm is a reliable machine learning technique that outperforms other methods in various aspects. Firstly, it is less prone to noise and overfitting than other models, thanks to its ensemble of decision trees that mitigates the variance of the final model. Secondly, it is computationally efficient and capable of processing large datasets with numerous features. Lastly, random forest allows for the determination of feature importance, which aids in the feature selection process. Overall, the random forest algorithm is a highly versatile and potent machine learning technique that can be utilized to address various problems. It is especially useful in situations that involve large datasets and high-dimensional features, and has the capability to provide valuable insights into feature importance. Random forest is an ensemble learn- ing approach that generates numerous decision trees in the training stage. The final output of the model is determined by consolidating the predictions of the individual trees, which can be either the mode of the

DOI: rp-9788770040723.140

classes for classification problems or the mean prediction for regression problems.

When utilizing the Random Forest (RF) Algorithm for regression-based issues, a fundamental measure used to evalu- ate the model's performance is the mean squared error (MSE).

The MSE aids in determining the degree to which data points are dispersed from each node of the decision tree. Specifically, it calculates the average of the squared differences between predicted and actual values of the response variable.A lower MSE implies better model performance.

On the other hand, when using the RF Algorithm for classification-based issues, the Gini-index is employed to determine how the decision tree branches. The Gini-index assesses the impurity or randomness of the class distribution of a node in the decision tree. A lower Gini-index suggests a less random distribution of classes and a more effective split. The RF Algorithm builds multiple decision trees using boot- strapped samples of the training data. Each tree is constructedby selecting the best split points based on the Gini-index.

Overall, the RF Algorithm is a versatile machine learning technique that can be applied to both regression and classification problems. Understanding which metrics to use and how they relate to the underlying problem can enhance the model's performance.

VIII. HYPERPARAMTER TUNING

Hyperparameter tuning is a widely-used technique in machine learning for enhancing a model's performance by optimizing its hyperparameters.Hyperparameters are predetermined before a model is trained and cannot be learned from data, such as regularization strength, learning rate, and the number of hidden layers.The primary objective of hyperparameter tuning is to identify the best combination of hyperparameters that result in the highest accuracy or lowest error rate on a validation dataset. This is typically accomplished by exhaustively searching through a range ofvalues for each hyperparameter and analyzing the model's performance on the validation dataset for each set of hyperparameters. Hyperparameter tuning can be executed using various approaches, such as random search, grid search, and Bayesian optimization. It is a crucial step in creating effective and precise machine learning models.In our case, after training the random forest model, we performed hyperparameter tuning to optimize its performance. Our results showed that the performance metrics improved significantly after hyperparameter tuning compared to the results obtained before tuning, as demonstrated in TABLE-II.Figure 4 shows a performance graph after hyperparameter tuning, compared to the graph in Figure 3 before hyperparameter tuning, clearly indicating a noticeable improvement in performance. Overall, hyperparameter tuning is a vital process for achieving the best performance in machine learning models.

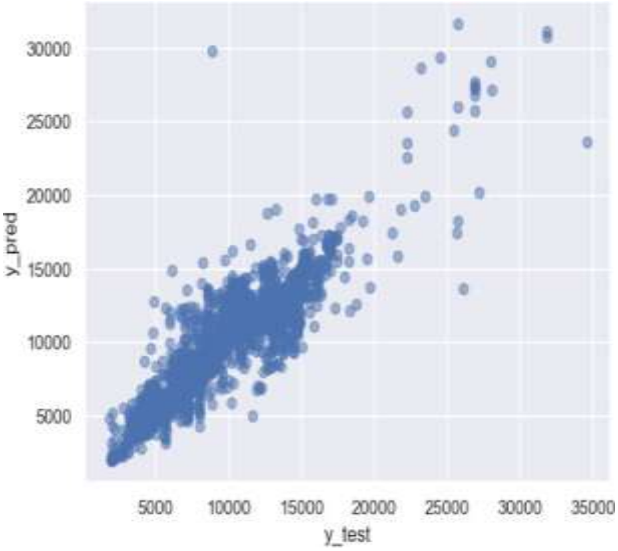

Fig. 3. Plot Performance Graph Before HP Tuning

Fig.4.PlotPerformanceGraphAfterHPTuning

TABLEII: PERFORMANCE METRICS BEFORE AND AFTER HP TUNING

Metrics	BeforeHP-Tuning	AfterHP-Tuning
R2	0.66	0.84
RMSE	1882.997	1773.43

IX. FUTURE SCOPE

In the future, our design will be able to accommodate datapertainingtothepurchaseofplanetickets.Thisinfowillbeabl etogivefurtherdetailsonacertainitinerary,suchasdepartureand arrivaltimesanddates,seatplacement,coveredauxiliarygoods,e tc.Bymergingdifferentformsofdataandworkingwithintherestr ictionsofthepresentsystem,itispossibletodevelopamodelthatis capableof delivering a more accurate and comprehensive hourly orevendailyforecastofairlinefares.Takingintoconsiderationof thecurrentindustrialdivisionandthemacroeconomicparameter s,this makes it feasible to estimate airline prices. Inaddition, the price of airfare in a given market sector may beaffected by an unanticipated rise in the number of passengerscaused

DOI: rp-9788770040723.140

by a particular special event. In order to complementourpredictionmodel,wewillcollectevent-relateddatafromarangeofresources,suchassocialmediaplatformsandnewsorganizations.Inaddition,wewillinvestigateothersophisticated machine learning models, such as deep learningtechniques, while concurrently striving to enhance the modelsalreadyinusebytweakingtheirhyper-parameterstocreatethemostefficientarchitectureforflightprice prediction.Prioritize the speed with which the model can anticipate theresultsoftheexperimentinadditiontothereliabilityofthemodel.Besidesthecurrentlyselectedfeatures,theremayexistotheraspectsthathavethepotentialtoimprovetheprecision of airfare price predictions. In the future, there is apossibility to extend this study to anticipate the airfare pricesfor an airline's entire flight network. However, this wouldrequire the utilization of a more extensive airfare dataset andconductingadditionalexperimentstovalidatethemodel'sperformance.

X. RESULTS

Throughout our project, we carried out both experimentaland research analyses on multiple machine learning models,includingK-Neighbours,DecisionTree,ExtraTree,andRandomForest.Afterconductingathoroughanalysisofthe obtained metrics, we concluded that the Random ForestModelperformedbetterthantheothermodels.InTABLEI,wehaveprovidedthescoresobtainedfromtrainingthedata with different ML models. Based on these results, wedevelopedamodelusingtheRandomForestalgorithm.Wethenperformedhyperparametertuning,whichfurtherimprovedtheperformanceofthemodel.Theperformancescores after hyperparameter tuning can be seen in TABLEII.Aftertrainingandevaluatingthemodel,wearethrilledtoreportthatitachievedanoutstandingaccuracyof95percentonthetrainingdataand82perconthetestdata.Thisoutcome is a testament to the model's high proficiency inprovidingprecisepredictionsofthetargetvariablewhenpresentedwithnewinputdata.Furthermore,ourresultshighlight the effectiveness of the Random Forest algorithm inresolving this specific problem and suggest its potential forsimilar tasks in the future.Overall, we are confident that ourresearch has made a significant contribution to the field andcould potentially be successfully implemented in real-worldscenarios.

XI. CONCLUSION

Theprimarypurposeoftheprojectwastoprovideassistancetousersinpredictingairfareand,consequently,inreducingtheexpensesassociatedwithbookingairlinetravel.Inordertoaccomplishthisgoal,asuitableMLtechniquemustbeselectedtotrainthemodel.Random-ForestAlgorithmhasbeenchosenasanidealalgorithmtouse for the project in order to obtain greater accuracy afterconductingtanin-depthstudy.Hyper-parameterTuningisdone after the Random Forest model training is completed, inorder to achieve an even higher level of precision and obtaintheverybestoutcomes.Furthermore,theprojecthighlightsthe importance of feature engineering and data preprocessing,asthesecansignificantlyimpacttheperformanceandaccuracyofthemodel.Nevertheless,thisinitialpilotstudyhashighlightedthepotentialofutilizingmachinelearningalgorithms to assist consumers in purchasing airfare ticketsduring thebestpossible market period.By taking advantageofthesemodels,consumerscanpotentiallysavemoneyandmake better-informed decisions when it comes to purchasingairfare.Overall,thisprojectdemonstratesthepowerofmachine learning in solving complex problems and providesvaluableinsightsandguidanceforfutureresearchinthefield.

REFERENCES

[1] K.DiamantarasandK.TziridisT.KalampokasG.Papakostas"Air-fare price prediction using machine learning techniques" in EuropeanSignalProcessingConference(EUSIPCO),DOI:10.23919/EUSIPCO.2017.8081365L.

[2] Supriya Rajankar, Omprakash rajankar and Neha sakhrakar "Flight farepredictionusingmachinelearningalgorithms"inInternationaljournalofEngineeringResearchandTechnology(IJERT)-June2019.

[3] T.Janssen, "A linear quantile mixed regression model for prediction of airline ticket prices".

[4] Samira Pouyanfar,Haiman Tian,Tianyi wang and YudongTao, A Framework for airline price prediction : A machine learning ap-proach".

[5] Neel Bhosale,Pranav Gole,Priti Lakade and GajananArsalwad, "Implementation Of Flight Fare Prediction System Using Machine Learning".

[6] Moshika, A., Thirumaran, M., Natarajan, B., Andal, K., Sambasivam, G., &Manoharan, R. (2021).Vulnerability assessment in heterogeneous web environment using probabilistic arithmetic automata. IEEE Access, 9, 74659-74673.

[7] Rajesh, M., &Sitharthan, R. (2022). Image fusion and enhancement based on energy of the pixel using Deep Convolutional Neural Network. Multimedia Tools and Applications, 81(1), 873-885.

[8] R.Ren,Y.Yang,andS.Yuan,"Prediction of airline ticket price,"UniversityofStanford,2014.