FLIGHT FARE PREDICTION

BSTRACT

In this paper, we present a review of customer side and airlines side prediction models. Our review analysis shows that models on both sides rely on limited set of features such as historical ticket price data, ticket purchase date and departure date. Features extracted from external factors such as social media data and search engine query are not considered. Therefore, we introduce and discuss the concept of using social media data for ticket/demand prediction. We used techniques like random forest regression , svm and exratree regressor for modelling the dataset. From the customer side, two kinds of models are proposed by different researchers to save money for customers: models that predict the optimal time to buy a ticket and models that predict the minimum ticket price

General Terms

Data Mining, Random Forest Regression, SVM

Keywords

Random Forest Regression, Extra Tree Regressor, Feature Selection

INTRODUCTION

Airline companies use complex algorithms to calculate flight prices given various conditions present at that particular time. These methods take financial, marketing, and various social factors into account to predict flight prices.

Nowadays, the number of people using flights has increased significantly. It is difficult for airlines to maintain prices since prices change dynamically due to different conditions. That’s why we will try to use machine learning to solve this problem. This can help airlines by predicting what prices they can maintain. It can also help customers to predict future flight prices and plan their journey accordingly.

The last two decades have seen steadily increasing research targeting both customers and airlines. Customer side researches focus on saving money for the customer while airline side studies are aimed at increasing the revenue of the airlines. Conducted researches employ a variety of techniques ranging from statistical techniques such as regression to different kinds of advanced data mining techniques.

It is possible that customers who bought a ticket earlier pay more than those who bought the same ticket later. Moreover, early

purchasing implies a risk of commitment to a specific schedule that may need to be changed usually for a fee

The objective of this project is to predict flight prices given the various parameters.

This will be a regression problem since the target or dependent variable is the price.

LITERATURE SURVEY

It is very difficult for the customer to purchase a flight ticket at the minimum price. For this several techniques are used to obtain the day at which the price of air ticket will be minimum. Most of these techniques are using sophisticated artificial intelligence(AI) research is known as Machine Learning.

Utilizing AI models, [2] connected PLSR(Partial Least Square Regression) model to acquire the greatest presentation to get the least cost of aircraft ticket buying, having 75.3% precision. Janssen [3] presented a direct quantile blended relapse model to anticipate air ticket costs for cheap tickets numerous prior days takeoff. Ren, Yuan, and Yang [4], contemplated the exhibition of Linear Regression (77.06% precision), Naive Bayes (73.06% exactness, Softmax Regression (76.84% precision) and SVM (80.6% exactness) models in anticipating air ticket costs. Papadakis [5] anticipated that the cost of the ticket drop later on, by accepting the issue as a grouping issue with the assistance of Ripple Down Rule Learner (74.5 % exactness.), Logistic Regression with 69.9% precision and Linear SVM with the (69.4% exactness) Machine Learning models.

Gini and Groves[2] took the Partial Least Square Regression(PLSR) for developing a model of predicting the best purchase time for flight tickets. The data was collected from major travel journey booking websites from 22 February 2011 to 23 June 2011. Additional data were also collected and are used to check the comparisons of the performances of the final model.

Janssen [3] built up an expectation model utilizing the Linear Quantile Blended Regression strategy for SanFrancisco to NewYork course with existing every day airfares given by www.infare.com. The model utilized two highlights including the number of days left until the takeoff date and whether the flight date is at the end of the week or weekday. The model predicts airfare well for the days that are a long way from the takeoff date, anyway for a considerable length of time close the takeoff date, the expectation isnt compelling.

Wohlfarth [15] proposed a ticket buying time enhancement model dependent on an extraordinary pre-preparing step known as macked point processors and information mining systems (arrangement and bunching) and measurable investigation strategy. This system is proposed to change over heterogeneous value arrangement information into added value arrangement direction that can be bolstered to unsupervised grouping calculation. The value direction is bunched into gathering dependent on comparative estimating conduct. Advancement model gauge the value change designs. A treebased order calculation used to choose the best coordinating group and afterward comparing the advancement model.

A study by Dominguez-Menchero [16] recommends the ideal buying time dependent on nonparametric isotonic relapse method for a particular course, carriers, and timeframe. The model gives the most extreme number of days before buying a flight ticket. two sorts of the variable are considered for the expectation. One is the passage and date of procurement.

Regression

Regression refers to a data mining technique that is used to predict the numeric values in a given data set. For example, regression might be used to predict the product or service cost or other variables. It is also used in various industries for business and marketing behavior, trend analysis, and financial forecast.

EXECUTION

- Importing Python Libraries and Loading our Data Set into a Data Frame.
- Splitting our Data Set Into Training Set and Test Set. ...
- Creating a Random Forest Regression Model and Fitting it to the Training Data. ...
- Visualizing the Random Forest Regression Results.

1.2 Random forest regressor

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Here we have applied Random Forest Regressor and we fitted X_train and Y_train Then we are predicting with respect to y _pred data. Finally we are finding our score for Training data and the test data Where the test data score is pretty close to Train data.

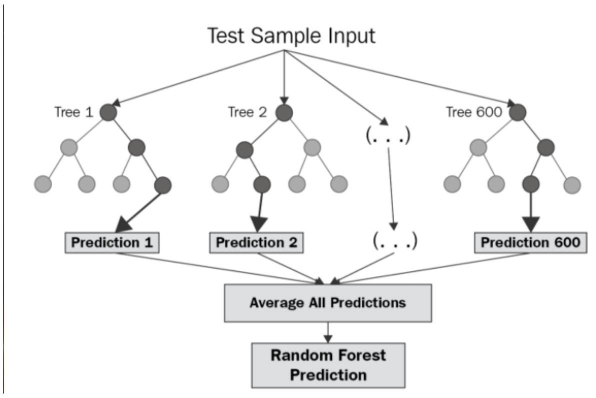


Figure 1

Random forest is a type of supervised learning algorithm that uses ensemble methods (bagging) to solve both regression and classification problems. The algorithm operates by constructing a multitude of decision trees at training time and outputting the mean/mode of prediction of the individual trees.

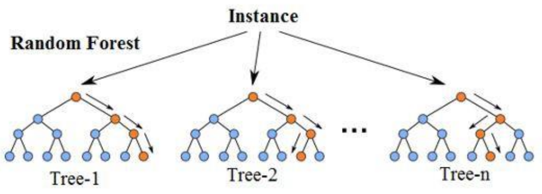
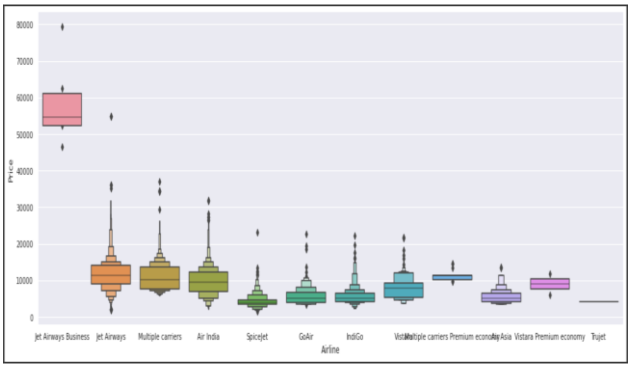
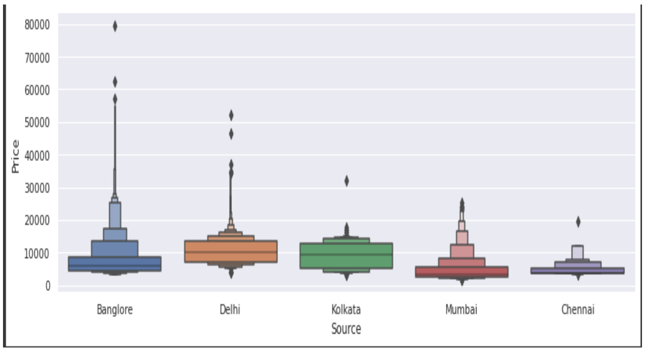


Figure 2

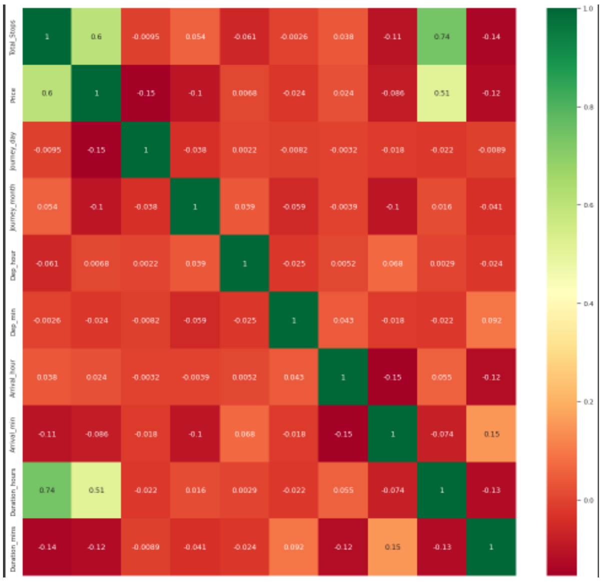
The fundamental concept behind random forest is the wisdom of crowds wherein a large number of uncorrelated models operating as a committee will outperform any of the individual constituent models. The reason behind this is the fact that the trees protect each other from their individual errors. Within a random forest, there is no interaction between the individual trees. A random forest acts as an estimator algorithm that aggregates the result of many decision trees and then outputs the most optimal result.



Handling Categorical Data



Source vs Price



Total_Stops	1	0.6	-0.0095	0.054	-0.061	-0.0026	0.038	-0.11	0.74	-0.14
Price	0.6	1	-0.15	-0.1	0.0068	-0.024	0.024	-0.086	0.51	-0.12
Journey_day	-0.0095	-0.15	1	-0.038	0.0022	-0.0082	-0.0032	-0.018	-0.022	-0.0089
Journey_month	0.054	-0.1	-0.038	1	0.039	-0.059	-0.0039	-0.1	0.016	-0.041
Dep_hour	-0.061	0.0068	0.0022	0.039	1	-0.025	0.0052	0.068	0.0029	-0.024
Dep_min	-0.0026	-0.024	-0.0082	-0.059	-0.025	1	0.043	-0.018	-0.022	0.092
Arrival_hour	0.038	0.024	-0.0032	-0.0039	0.0052	0.043	1	-0.15	0.055	-0.12
Arrival_min	-0.11	-0.086	-0.018	-0.1	0.068	-0.018	-0.15	1	-0.074	0.15
Duration_hours	0.74	0.51	-0.022	0.016	0.0029	-0.022	0.055	-0.074	1	-0.13
Duration_mins	-0.14	-0.12	-0.0089	-0.041	-0.024	0.092	-0.12	0.15	-0.13	1

Correlation Heatmap

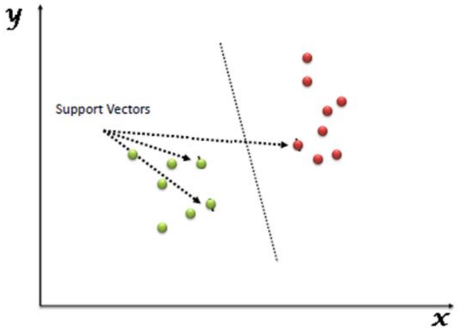
1.3 SVM - Support Vector Machine

In the proposed paper Support Vector Machine used as regression analysis that relays on kernel function considered as non parametric technique. The following kernels are used: Linear, Polynomial, Radial Basis Function.

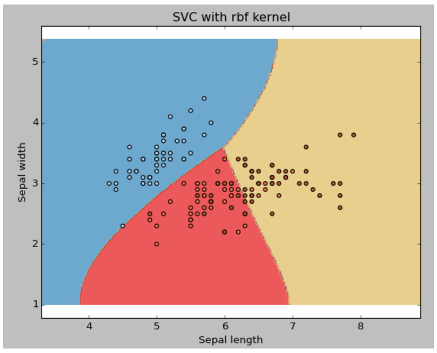
As per the previous studies Random forest and the gradient boosting gives the maximum accuracy OF 79.6%

SVM algorithm gives 80.6% of accuracy.

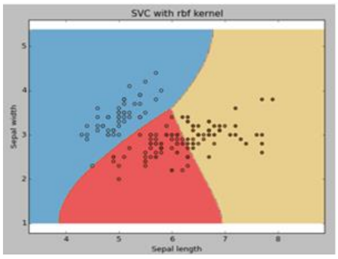
"Support Vector Machine" (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate.



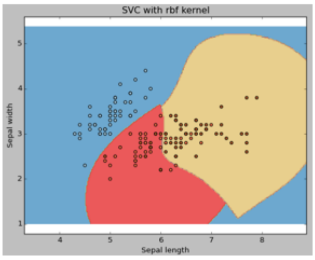
In Python, scikit-learn is a widely used library for implementing machine learning algorithms. SVM is also available in the scikit-learn library and we follow the same structure for using it(Import library, object creation, fitting model and prediction).



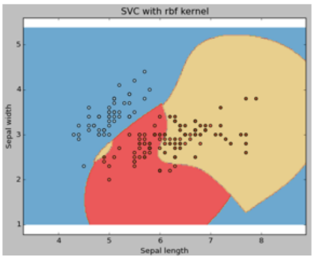
c =1



C =100

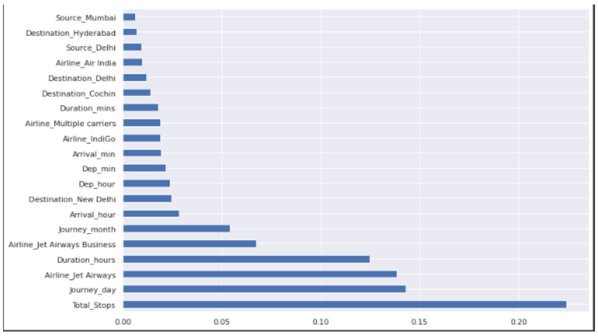


c =1000



1.4 Extra tree regressor

The Extra Trees algorithm works by creating a large number of unpruned decision trees from the training dataset. Predictions are made by averaging the prediction of the decision trees in the case of regression or using majority voting in the case of classification.



Importing feature using ExtraTreesRegressor

COMPARATIVE ANALYSES

Table 1 Random Forest and SVM Algorithms

ML algorithms	R-squared	MAE	MSE
Random forest	0.67	0.08	0.04
SVM	-.12	0.21	0.11

EXPERIMENTAL ANALYSES

Proposed Methodology

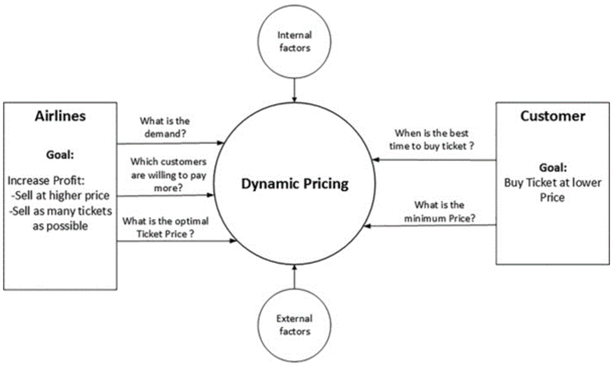
Flight fare prediction

A significant number of research works exits that proposed prediction models for dynamic pricing in airlines which can be classified into two groups: demand prediction .

Early prediction of the demand along a given route could help an airline company preplan the flights and determine appropriate pricing for the route. Existing demand prediction models generally try to predict passenger demand for a single flight/route and market share of an individual airline. Price discrimination allows an airline company to categorize customers based on their willingness to pay and thus charge them different prices. Customers could be categorized into different groups based on various criteria such as business vs leisure, tourist vs normal traveler, profession etc. For example, business customers are willing to pay more as compared to leisure customers as they rather focus on service quality than price.

Despite the fact that there are several studies conducted on both sides, customer and airlines, no attempt has been made to present a literature survey and review of existing work.

Therefore, the main goal is to present a comprehensive literature review of existing studies related to this topic which can be utilized by future researchers. We first classify and present existing studies into two categories based on their desired goal (customer side models and airline side models). We then group existing work based on the specific problem being addressed. Several issues have been discussed including data sources, features and various techniques employed for prediction.

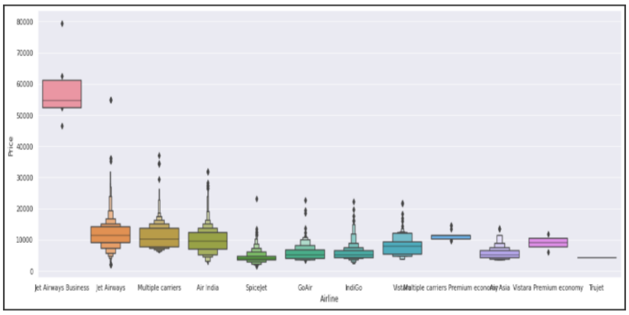

Internal factors
Airlines
Goal:
Increase Profit:
-Sell at higher price
-Sell as many tickets as possible
What is the demand?
Which customers are willing to pay more?
What is the optimal Ticket Price ?
Dynamic Pricing
When is the best time to buy ticket ?
What is the minimum Price?
Customer
Goal:
Buy Ticket at lower Price
External factors

Flight fare prediction

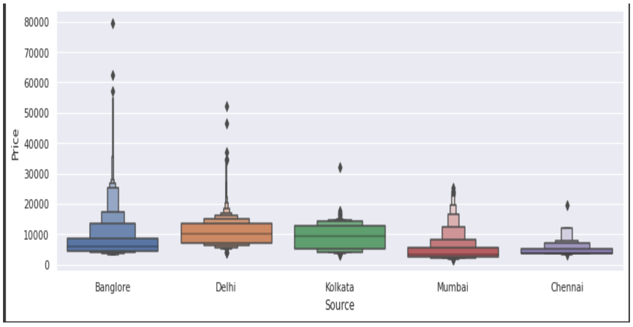
Once the model is trained, we have UI kind off a web app where the user will be asked to enter the departure date and time and the arrival date and time. And they will be asked to enter the source and destination, and the stoppage and finally they are allowed to choose the airline.

After this process they will be asked to submit doing by which the flight fare for above mentioned details will be shared to the user effectively.

Visualization in graph



Handling Categorical Data



Source vs Price

CONCLUSION

We were successfully able to analyse each route and generalize the entire project based in terms of the sector to which the route belonged, and classified them into three major subsections - Business Routes, Tourist Routes and Tier-2 Routes.

We have also successfully busted some of the typical myths and misconceptions related to the airline industry and backed them up with data and analysis.

Finally, we have created a User Interface for the entire process of buying an airline ticket and given a proof of our predictions based on the previous trends with our prediction. Thus leaving it as a battle between 'The risk appetite of the user' vs 'Our understanding of the airline industry'.

Future Work

- ➢ More routes can be added and the same analysis can be expanded to major airports and travel routes in India.
- ➢ The analysis can be done by increasing the data points and increasing the historical data used. That will train the model better giving better accuracies and more savings.
- ➢ More rules can be added in the Rule based learning based on our understanding of the industry, also incorporating the offer periods given by the airlines.
- ➢ Developing a more user friendly interface for various routes giving more flexibility to the users.

REFERENCES

[1] O. Etzioni, R. Tuchinda, C. A. Knoblock, and A. Yates. To buy or not to buy: mining airfare data to minimize ticket purchase price.

[2] Manolis Papadakis. Predicting Airfare Prices.

[3] Groves and Gini, 2011. A Regression Model For Predicting Optimal Purchase Timing For Airline Tickets.

[4] Modeling of United States Airline Fares – Using the Official Airline Guide (OAG) and Airline Origin and Destination Survey (DB1B), Krishna Rama-Murthy, 2006.

[5] B. S. Everitt: The Cambridge Dictionary of Statistics, Cambridge University Press, Cambridge (3rd edition, 2006). ISBN 0-521-69027-7