EAS 508: Project Preliminary Report

Group 31: Sai Teja Dondeti, Sai Teja Sankarneni, Nishant Varma Indukuri, Subba Rao Koduri

Title: Flight Fare Prediction Using Machine Learning

Abstract:

Airline companies offer air transport services, Airline companies may form alliances with other charter companies for offering these services in which they operate and share the same flight. Airline companies assign prices to their tickets to maximize profits. There are many factors that influence the flight tickets such as flight duration, days left for departure, arrival time and departure time etc. Each factor has its own affect based on its algorithm.

Introduction

The objective of the study is to analyze the flight fare prediction dataset and to conduct various statistical hypothesis tests to get meaningful information from it. With this dataset, we will predict the flight ticket price. The price may rely upon different factors. Each factor has its own proprietary rules and algorithms to set the price accordingly. We will use the advances in Machine Learning (ML) to infer such rules and model the price variation.

Data

We have extracted data from EaseMyTrip.com via Kaggle. EaseMyTrip is an Indian online travel company, that provides hotel bookings, air tickets, holiday packages, bus bookings, and white-label services.

Data was collected in two parts: one for economy class tickets and another for business class tickets.

A total of 300153 unique flight booking options were extracted. The data was than normalized into 5 dimension and 1 fact tables.

Airline(AirlineID, Airline) - This table consists of distinct airlines for which the data is collected,

Arrival_Time(ArrivaltimeID, scheduled_arrival) - This table consists of distinct part of the days during which the flight either arrives or departs from the airport.

City(CityID, City) - This table have 6 cities between the journey is collected.

Flights(FlightID, Flight_Number, source_city, destination_city) - Flights have the distinct flight information with flight number and its source and destination cities.

Number_of_stops(Stop_id, Stop_count) - This have the information on number of stops airline have from source to destination cities.(one, two_or_more, zero)

Ticket_class(TicketID, ticket_type) - This Table have information about type of ticket(business, economy)

Flight_Fare(Flight_Fare_ID, airlineID, FlightID, departure_time_id, stop_id, arrival_time_id, ticketID, duration, days_left, price) - This is the normalized table that have information about all the bookings that we have collected from source files.

analysis

We have analyzed the various factors that affect the ticket price such as Airline (Vistara have highest ticket price,and air asia accounts for the least), Ticket class (Business class have higher rates and economy have least ticket), flights with single stop have more price and some other factors are analyzed and considered.

Model

After normalizing the data and performing various regression techniques we have finalized that ExtraTreesRegressor is the best regression technique for the data we have obtained. The accuracy of each model is

	Model_Name	Adj_R_Square
0	BaggingRegressor	0.984874
1	ExtraTreesRegressor	0.984091
2	RandomForestRegressor	0.982971
3	LinearRegression	0.904653
4	Ridge Regression	0.904653
5	Lasso Regression	0.904653

From the Adj_R_Square values we have finalized that ExtraTreesRegessor is the best model for out data.

Conclusion

From the above observations and with different Regression techniques used, Bagging Regressor and ExtraTreesRegressor provided best predictions with good accuracy and are more suitable models for the above analysis.

UB Box Link

References