PREDICTION OF AIRFARE USING MACHINE LEARNING

Name: Mr. Tushar Singh Email: tushar.24.ts@gmail.com

Graduate, EXTC ,Ramrao Adik Institute of Technology, Navi Mumbai

ABSTRACT: Predicting approximate air fare at a particular time has become one of the utmost important tasks. People who travel more often usually have an idea when the fares will be high and when the fares will be low. Prices are usually high when the demand is high to limit the capacity and vice versa. In this paper will try to extract the air fare of one specific route and using exploratory data analysis and ML models will try to find the trends as well as approximate prediction of the air fare.

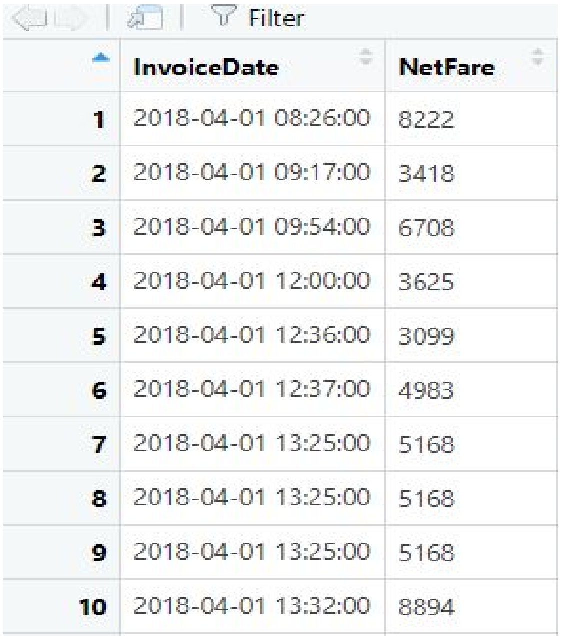
1. INTRODUCTION

It's common practice that people try to book tickets at least 1-2 weeks prior to the day of flight. The change in the air fare is mainly controlled by the air agencies and they have the luxury to change it accordingly which benefits their revenue. Whenever the demand is high the prices are sky rocket high to maximize their profit. There are various algorithms which are not in the public domain that a common man can study and analyze. With the use of Machine learning, Artificial Intelligence and statistics we can try to predict the air fare. Regression models usually help in drawing out a relationship between dependent and independent variables. Major issue in this due process is difficulty in procuring the proper data as there are very less accurate websites and sources available which provide the data. Will be using various publicly available datasets to predict the air fare. Customers usually try to book tickets in advance in order to avoid the high fare, this has been one of the common techniques which customers adopt in order to benefit themselves.

2. DATA COLLECTION

There are various websites and datasets available consisting of air fare and other factors associated it with. The dataset used for this article was provided by a startup company named Innodatatics.

2.1 DATA DESCRIPTION

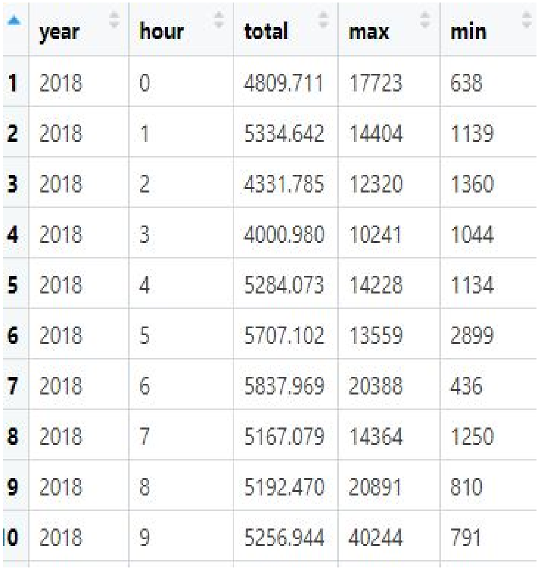


	InvoiceDate	NetFare
1	2018-04-01 08:26:00	8222
2	2018-04-01 09:17:00	3418
3	2018-04-01 09:54:00	6708
4	2018-04-01 12:00:00	3625
5	2018-04-01 12:36:00	3099
6	2018-04-01 12:37:00	4983
7	2018-04-01 13:25:00	5168
8	2018-04-01 13:25:00	5168
9	2018-04-01 13:25:00	5168
10	2018-04-01 13:32:00	8894

So, this data was collected during various intervals in a 24-hour clock for a period of 365 days. Invoice date is the time and date at which the transaction took place to purchase the ticket and Net fare is the amount paid to the agency. In this article our aim is to study the air fare of one single trip only during various intervals in order to draw conclusions which time is the most expensive and which time is the cheapest to book the flight.

2.2 CLEANING AND PREPARING DATA

Procuring the data was not much of a big task as it was readily available but the cleaning process took most of the time. Invalid data such as negative values, misprinted values and outliers were removed along with duplicate values. Features such as day of the week, avg fare for the month are also considered for determining various trends.



	year	hour	total	max	min
1	2018	0	4809.711	17723	638
2	2018	1	5334.642	14404	1139
3	2018	2	4331.785	12320	1360
4	2018	3	4000.980	10241	1044
5	2018	4	5284.073	14228	1134
6	2018	5	5707.102	13559	2899
7	2018	6	5837.969	20388	436
8	2018	7	5167.079	14364	1250
9	2018	8	5192.470	20891	810
10	2018	9	5256.944	40244	791

3. MACHINE LEARNING MODELS

Will be using various forecasting techniques such as simple exponential, ARIMA, ETS, Linear Regression. Packages like forecast, xts, nnetar were used. The parameters like RMSE, MAE and MSE are considered to verify the performance of these models.

3.1 LINEAR REGRESSION

To determine the correlation between two continuous variables, simple linear regression analysis is used. One of the two variables is the predictor variable of which value is to be found. It gives the statistical relationship not the deterministic relationship between two variables. Linear regression algorithm gives the best fit line to the given data for which the prediction error is minimum. Gradient descent and cost function are the two major factors to understand linear regression.

The equation for linear regression is: y(pred) = b0+b1 ∗ x (1)

The value of coefficients b1 and b0 are chosen so that the error value is as small as possible. The square of predicted and actual value difference gives the error. To deal with the negative values, the mean square error is taken (MSE). Here b0 gives the positive or negative relationship between the x and y, whereas b1 is called bias. The accuracy of the regression problem is measured in terms of R-squared, MAE, RMSE.

3.2 NAIVE MODEL

This is one of the simplest forecasting techniques in which the forecasted values are equal to the previous values. The set is prepared without adjusting any value, simply the previous value is being projected as the future value.

The equation for naïve forecasting is $\text{^}y_{T+h|T}=y_T$.

Where ^Yt+h is the previous value and Yt is the present value.

3.3 SIMPLE EXPONENTIAL SMOOTHING

This is one of the simplest exponential smoothing techniques. In naïve method which only considers the last value and all other values have no importance but in the case of Simple exponential smoothing importance is also given to previous values and their importance is decreasing exponentially. Forecasts are calculated using weighted averages, where the weights decrease exponentially as observations come from further in the past, the smallest weights are associated with the oldest observations:

$$^\wedge y_{T+1|T}=\alpha y_{T}+\alpha(1-\alpha)y_{T-1}+\alpha(1-\alpha)_{2}$$

The values of alpha vary from 0 to 1.

3.4 ETS

ETS stands for Error, Trend and Seasonality. Various values are assigned to (E, T, S) in order to get a perfect model.

A= Additive model

M= Multiplicative

N= None

Z= Automatically select

Ets models are mostly used in case where:

1. Data is not stationary
2. use exponential smoothing
3. use if there is a trend and/or seasonality in the data, as this model explicitly models these components

3.5 ARIMA

Unlike previous methods which consider trend, seasonality Arima model is generally used for stationary data. The best description for stationary data is one which does not have a trend or seasonality.

Below are various graphs consisting of stationary and non-stationary time series data.

Differencing: This technique is used to normalize the trend or seasonality present in the data. Consecutive values are subtracted from each other in order to keep the data around the mean.

Arima model: This model takes three values as an input (p, d, q):

Where p and q values are used to specify the number of significant lags to be considered for the correlation and moving average altogether is considered in the same model.

“d” stands for number of times differencing is required to normalize the data around the mean.

4.EXPERIMENTAL RESULTS:

Various models like ets, ses, arima were used to construct models and then predict future values. These predicted values are then compared with the test data and the model having least RMSE value is the best model.

Model Name	RMSE VALUE
Linear Regression	275.65
Simple Exponential smoothing	235.68
ETS	255.32
Naïve	198.35
ARIMA	193.35

5. CONCLUSION: To evaluate the conventional algorithm, a dataset is built for route BOMBAY to DELHI and studied a trend of price variation for the period of limited days. Machine Learning algorithms are applied on the dataset to predict the dynamic fare of flights. RMSE is the measure which is used to compare various models and it turns out Arima and Naïve methods had the lowest RMSE value. Also, we conclude that even with limited data it was possible to predict air fare with more data it would only increase the accuracy of the prediction.

6. FUTURE WORK:

With the help limited data one can only predict few trends in the data for e.g.: weekday flights were less expensive than weekend flights, non-seasonal months had cheap rates compared to months with more holidays. Hence, to improve the accuracy one needs to collect a wide range of data.