

**PREDICTIVE ANALYTICS**

***Topic : Flight Fare Prediction***

*Adithya Vinod*

*Sneha Kumble Manohara*

**School of Graduate Professional Studies**

MPS in Data Analytics

IE 575 – PREDICTIVE ANALYTICS

(Term3, 2023)

# Document Control

## Work carried out by:

|  |  |  |
| --- | --- | --- |
| **Name** | **Email Address** | **Task description** |
| Adithya Vinod | avv5502@psu.edu |  |
| Sneha Kumble Manohara | skm6622@psu.edu |  |
|  |  |  |

## Revision Sheet

|  |  |  |
| --- | --- | --- |
| **Release No.** | **Date** | **Revision Description** |
| 1 | 01/15/2023 | Research Goal and Business Understanding |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**TABLE OF CONTENTS**

Contents

[Document Control 1](#_Toc124682087)

[Work carried out by: 1](#_Toc124682088)

[Revision Sheet 1](#_Toc124682089)

[Week 1 Assignment 4](#_Toc89640385)

[Research Goals and Business Understanding 4](#_Toc89640386)

**General Guidelines**

1. Please note that it is a team-based project. Each team can have no more than 2 members.
2. To complete all the assignments in this course please use this template document.
3. Each assignment has to be submitting by the following Sunday 11:59 PM EST.
4. Each figure should be followed by a brief description about the figure.
5. The figures can be hand drawn and scanned in some circumstances, but the hand drawn figure should be clear and legible to obtain full credits. Unclear hand drawn figures will receive partial credits. For drawing figures and diagrams, it is advised to use tools.
6. Figures and tables should have appropriate captions. For documenting and referencing styles please follow the APA or MLA writing style.
7. Please make sure that you provide a reference section.
8. Any material text or figure taken from books, journals or Internet should be referenced. If you have a sentence or a figure that does not belong (authorship) to you, they need to be clearly referenced. If you fail to do so your report will be considered as a case for plagiarism. It is your duty to make sure that your report is free from any activity related to plagiarism. In case you are suspected of attempting plagiarism then you will be responsible for the cause. The penalty for plagiarism will be a “0” awarded to your report. So, it is good to keep simple, always have the principle to acknowledge people for their contributions.

Please go through the following instructions before submitting the report

#### **Academic Integrity**

Academic integrity — scholarship free of fraud and deception — is an important educational objective of Penn State. Academic dishonesty can lead to a failing grade or referral to the [Office of Student Conduct](http://www.sa.psu.edu/ja/).

Academic dishonesty includes, but is not limited to:

* cheating
* plagiarism
* fabrication of information or citations
* facilitating acts of academic dishonesty by others
* unauthorized prior possession of examinations
* submitting the work of another person or work previously used without informing the instructor and securing written approval
* tampering with the academic work of other students

#### How Academic Integrity Violations Are Handled

In cases where academic integrity is questioned, [procedure requires an instructor to notify a student](http://www.psu.edu/oue/aappm/G-9-academic-integrity.html) of suspected dishonesty before filing a charge and recommended sanction with the college. Procedures allow a student to accept or contest a charge. If a student chooses to contest a charge, the case will then be managed by the respective college or campus Academic Integrity Committee. If a disciplinary sanction also is recommended, the case will be referred to the [Office of Student Conduct](http://www.sa.psu.edu/ja/title=).

All Penn State colleges abide by this Penn State policy, but review procedures may vary by college when academic dishonesty is suspected. Information about Penn State's academic integrity policy and college review procedures is included in the information that students receive upon enrolling in a course.

Additionally, Penn State students are expected to act with civility and personal integrity; respect other students' dignity, rights, and property; and help create and maintain an environment in which all can succeed through the fruits of their own efforts. An environment of academic integrity is requisite to respect for oneself and others, and a civil community.

#### For More Information on Academic Integrity at Penn State

Please see the [Academic Integrity Chart](http://www.campuses.psu.edu/CAO.pdf)  for specific college contact information or visit one of the following URLs:

* Penn State Senate [Policy on Academic Integrity](http://www.psu.edu/dept/oue/aappm/G-9.html)
* [iStudy for Success!](http://istudy.psu.edu/tutorials/) — learn about plagiarism, copyright, and academic integrity through an educational module
* [Turnitin](http://tlt.its.psu.edu/turnitin) a web-based plagiarism detection and prevention system

**Week 1 assignment**

**Research Goal:**

The airline industry is considered as one of the most sophisticated industries in using complex pricing strategies. Now-a-days flight prices are quite unpredictable. The ticket prices change frequently. Customers are seeking to get the lowest price for their ticket, while airline companies are trying to keep their overall revenue as high as possible. Airlines use sophisticated tactics, which they call "revenue management". Optimal timing for airline ticket purchasing from the consumer’s perspective is challenging principally because buyers have insufficient information for reasoning about future price movements.

Using Predictive Analytics, it is possible to reduce the uncertainty of flight prices. So here we will be predicting the flight prices using efficient machine learning techniques.

**Business questions:**

* 1. Prediction of flight price.
  2. Which Source has the best price?
  3. Which destination has the best price
  4. Which airline is cheaper?

**Dataset:**

The dataset for the above problem statement can be found on Kaggle. The link to the dataset is:

<https://www.kaggle.com/datasets/nikhilmittal/flight-fare-prediction-mh>

**Week 2 assignment**

In week 1, you had defined your research/business goal (queries) and had also identified the source(s) for data collection. In this week your task will be to retrieve/collect the data from the source and maintain a local copy of the dataset/database/ data mart/data warehouse. To pursue a challenging as well as a manageable project I will recommend that your dataset/database/ data mart/data warehouse contains at least 20 +, 30 +, < 40 columns and several thousand rows (not more than 100K unless you need to have a larger dataset). Also note that you can establish a dataset/database/ data mart/data warehouse by combining one or more dataset/database/ data mart/data warehouse.

**Note**: In this course you are required to work only with structured data.

**For structured data**

Once you have retrieved/collected the dataset/database/data mart/data warehouse update the template document with the following sections:

* Provide a brief description of the dataset/database/ data mart/data warehouse. The description could involve listing the following: the attributes, the data types of the attributes, number of instances, attribute values, missing values, outliers if any? etc. You can follow the format for the data description provided in the link <http://archive.ics.uci.edu/ml/machine-learning-databases/car/car.names.>
* Also describe the types of data and measurement scales (nominal, ordinal, interval, and ratio) across each attribute. Please go through this link (<http://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/>) for additional details about the types of data and measurement scale.

Submit your completed week 2 assignment to the dropbox in Canvas.

**Week 2 assignment**

The dataset was downloaded from Kaggle. It consists of 11 columns with data types including categorical, numeric, date and time.

We have 2 datasets here — training set and test set.

The training set contains the features, along with the prices of the flights. It contains 10683 records, 10 input features and 1 output column — ‘Price’.

The test set contains 2671 records and 10 input features. The output ‘Price’ column needs to be predicted in this set.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Data Type** | **Data Classification** | **Description** |
| **Airline** | Categorical | Nominal | Airline Name e.g. IndiGo, Air India |
| **Date\_of\_Journey** | Date Time | Time Series | Date of journey |
| **Source** | Categorical | Nominal | Source of travel |
| **Destination** | Categorical | Nominal | Travel destination |
| **Route** | Categorical | Nominal | Journey route e.g. BLR → DEL |
| **Dep\_Time** | Date Time | Time Series | Departure time |
| **Arrival\_Time** | Date Time | Time Series | Arrival time |
| **Duration** | Date Time | Time Series | Travel duration |
| **Total\_Stops** | Categorical | Nominal | Total number of stops e.g. 2 stops, non-stop |
| **Additional\_Info** | Categorical | Nominal | Additional Information e.g. In-flight meal not included |
| **Price** | Numeric | Interval | Flight fare |

**Week 3 assignment**

As a part of data cleaning, we found only one row that has NA, which we decided to drop.

Table

Description automatically generated

**Creating new features:**

There we few columns that needed preprocessing by creating new columns like day, month and year of journey from Date\_of\_Journey, Departure minute and hour from Departure\_time and Arrival minute and hour from Arrival\_time.

|  |  |
| --- | --- |
| **Before** | **After** |
| Date\_of\_Journey | day\_of\_journey, month\_of\_journey, year\_of\_journey |
| Dep\_Time | Dep\_Time\_hour, Dep\_Time\_minute |
| Arrival\_Time | Arrival\_Time\_hour, Arrival\_Time\_minute |
| Duration | Duration\_hour, Duration\_minute |

**Feature Removal:**

Variables that wouldn’t be useful for predicting the Flight fare were dropped. The variables included:

* 1. year\_of\_journey
  2. Dep\_Time
  3. Arrival\_Time
  4. Date\_of\_Journey
  5. Duration

**Outliers:**

Price and Duration\_hour column had outliers which were detected using the box and whisker plot. Instead of removing outliers, we will be using Quantile based flooring and capping method, where the data points that are lesser than the 10th percentile are replaced with the 10th percentile value and the data points that are greater than the 90th percentile are replaced with 90th percentile value.

Chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

**Skewness:**

We did not notice a lot of skewness in the data. To reduce the skewness, we will be implementing box-cox transformation,

Skewness before transformation

Text

Description automatically generated

Chart

Description automatically generated

**Week 4 assignment**

As a part of data cleaning, there were few columns that needed preprocessing by creating new columns like day, month, and year of journey from Date\_of\_Journey, Departure minute and hour from Departure\_time and Arrival minute and hour from Arrival\_time. Variables that wouldn’t be useful for predicting the Flight fare were dropped. We performed boxcox transformation, to remove the skewness of the data. Price and Duration\_hour column had outliers.

As a part of data cleaning, we found only one row that has NA, which we decided to drop.

Table

Description automatically generated

**Outliers:**

In order to remove the outliers, we used Quantile based flooring and capping, where the outlier is capped at a certain value above the 99th percentile value or floored at a factor below the 1st percentile value.

Before outlier removal: After outlier removal:

Chart

Description automatically generatedChart, histogram

Description automatically generated

Before outlier removal: After outlier removal:

Chart, box and whisker chart

Description automatically generated**Chart, histogram, box and whisker chart

Description automatically generated**

**Skewness:**

We have considered a feature to be skewed if the skewness value lies outside the range: [-0.5, 0.5].

The following features were skewed, namely:

1. Price
2. Duration\_hour

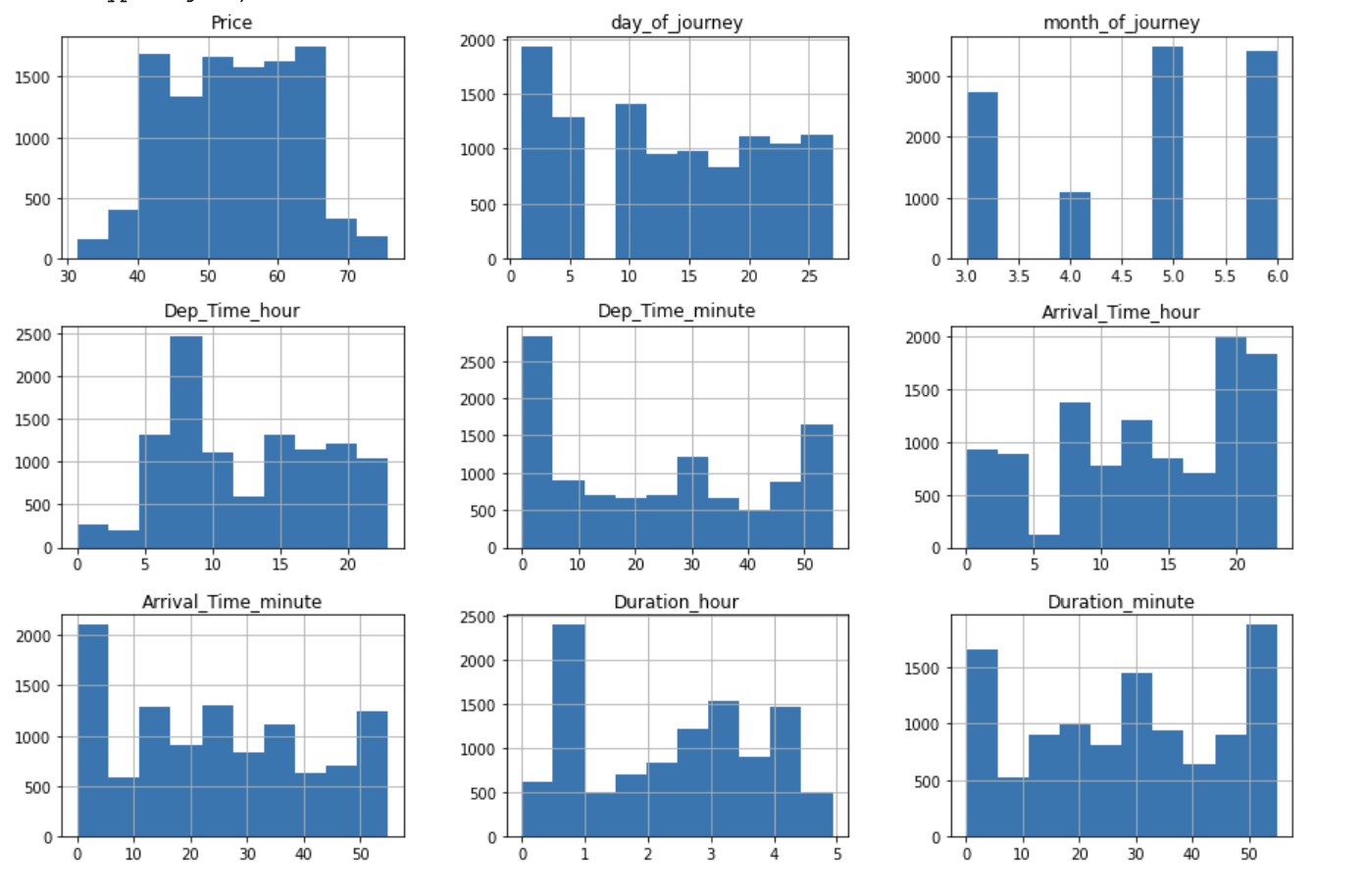
Boxcox transformation was implemented to reduce the skewness in the data.

**Skewness before transformation:** **Skewness after transformation:**

**Text

Description automatically generated Text

Description automatically generated**

****

**Feature Scaling:**To improve the optimization process by making the flow of gradient descent smoother and helping algorithms reach the minimum of the cost function more quickly, we have scaled the dataset. We have used the Min-Max scaler to scale the data to the range [0,1].

**A picture containing text, crossword puzzle, vector graphics

Description automatically generated**

**Week 5 assignment**

We performed several steps like variable selection, transformation, data reduction etc. As a part of preprocessing, we created new features from the existing ones and removed some of the features that were not necessary for the prediction of flight price.

To reduce the outlier, we used Quantile based flooring and capping, where the outlier is capped at a certain value above the 99th percentile value or floored at a factor below the 1st percentile value. Boxcox transformation was implemented to reduce the skewness in the data.

A graph was plotted In-order to answer some of the business queries like the average price in each source and destination, average ticket price for each airline and average price for different number of stops. We performed stratified sampling based on airlines. The dataset is then split into Train and test in the ratio 70:30.

We will be calculating Variance inflation factor(VIF) to check the multicollinearity. Since there are variables that are one-hot encoded, these variables will have VIF value of inf. We will be considering just the numerical variables for feature reduction. After plotting the Correlation plot, we can visualise that there is a very strong correlation between Total\_stops and Duration\_hour. Hence, one of them needs to be dropped. On calculating the VIF scores for the numerical data, we can see that there are a few varaibles that have a VIF score greater than 5, indicating autocorrelation. We try and drop Total\_stop and month\_of\_journey. We will re-calculate the VIF to see if the autocorrelation has reduced. Since all the VIF values are less than 5, we can say that that the autocorrelation has reduced.

Further, we will be considering features that have feature importance of atleast 0.5% for the prediction of flight price.

**Predictive analytics techniques**

We plan to use regression analysis technique to predict the flight price. Regression analysis is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. There are few Tree based regressors and some non-tree based regressor like clustering models(KNN) and linear model(Logistic, SVR). We will be considering all of them for prediction based on their performance. The algorithms that perform well among them will be considered for hypertunning. Some of the regression models that we will be using are Gradient boosting regressor, Random Forest, LGBM Regressor and Bagging Regressor. Initially we will evaluate these regression models without any hyperparameter through Cross Validation with 10 folds.

There are several performance measures that are commonly used to evaluate the accuracy of regression models:

1. Mean Absolute Error (MAE): The average absolute difference between the predicted values and the true values.
2. Mean Squared Error (MSE): The average of the squared differences between the predicted values and the true values.
3. Root Mean Squared Error (RMSE): The square root of the mean squared error, which provides a more interpretable result in the same units as the response variable.
4. R-squared (R2): A measure of how well the model fits the data, with a value close to 1 indicating a good fit.
5. Mean Absolute Percentage Error (MAPE): The average absolute difference between the predicted values and the true values, expressed as a percentage of the true values.

The metric chosen for evaluation will be the R-square, a statistical measure that represents the proportion of the variance of a dependent variable that is explained by an independent variable or variables in regression problems. Thus, the larger the R², the more explanatory the linear model is, that is, the better it fits the sample.

**Week 6 assignment**

The prediction of flight prices is important because it can help travelers plan their trips and make informed decisions about when to book their flights. By knowing the likely price of a flight, travelers can budget accordingly and avoid paying more than they need to for their travel. Flight prices can be highly variable and affected by a range of factors, such as supply and demand, seasonal changes, airline promotions, and global events. Predicting these fluctuations in advance can help travelers make decisions that are both economical and convenient.

Using predictive analytics can be very helpful in predicting flight prices. Predictive analytics is a process that involves using statistical models and machine learning algorithms to analyze data and make predictions about future events or trends. When it comes to predicting flight prices, predictive analytics can help by identifying patterns and trends in historical data that are indicative of future price changes.

Below is the detailed explanation of our findings.

1. **Prediction of flight price.**

We used regression analysis technique to predict the flight price. First, we stratified the sample based on the airlines. Then checked for multicollinearity based on VIF. Since there were variables that were one-hot encoded, these variables had VIF value of inf. Hence, we considered just the numerical variables for feature reduction. On calculating the VIF scores for the numerical data, we could see that there were a few variables that had a VIF score greater than 5, indicating autocorrelation. After dropping Total\_stop and month\_of\_journey and recalculating VIF, We were able to see that all the VIF values were less than 5, and hence say that the autocorrelation was reduced. After checking variable importance, 20 variables with importance of at least 0.5 were considered for the prediction of the price.

Chart, histogram

Description automatically generated

There are few Tree based regressors and some non-tree based regressor like clustering models (KNN) and linear model (Logistic, SVR). We considered all of them for prediction of the price based on their performance. The algorithms that performed well among them were considered for hyper tuning. Some of the regression models that were used are Gradient boosting regressor, Random Forest, LGBM Regressor and Bagging Regressor. Initially we evaluated these regression models without any hyperparameter through Cross Validation.

Graphical user interface, text

Description automatically generated

We tested non-tree-based algorithms like KNeighborsRegressor, LinearRegression and SVR on normalized and standardized data and obtained an improvement in the results in both cases. However, tree-based models still showed the best results.

**Text

Description automatically generated**

Further, we chose the best performing models evaluated through cross validation to apply hyper tuning. After applying hyper tuning and evaluating the models, LGBMRegressor showed the best fit based on the R-squared value obtained.

**Graphical user interface, text

Description automatically generated**

1. **Which Source has the best price?**

The dataset we chose had the details of different Indian airports as sources and destinations. This data needed to be cleaned and normalized. Once this process was completed, performing exploratory data analytics would open doors to new insights. Not all sources have the same number of tickets booked, hence a mean value of ticket prices is plotted against the various Indian airport sources.

On plotting a bar chart, we understand that Chennai and Mumbai seem to be the destinations with the least flight fare, whereas Delhi and Kolkata are comparatively costlier. This information can be used to make data-driven decisions, such as choosing a more cost-effective airport to travel from or planning a trip to a destination with lower flight fares.

**Chart, funnel chart

Description automatically generated**

1. **Which destination has the best price?**

Once we realized that not all source airports have the same ticket prices, it was interesting to understand if destination prices varies as well. A similar approach as above was taken by plotting the graph of destination vs mean price. We were able to find the source that had the best price.

Hyderabad and Kolkata were the sources that had cheaper flight fare.   
Whereas Cochin and Bangalore were costlier destination. This information can be useful for people who are planning to travel and are looking for cost-effective options.

While there may be other factors besides the source and destination airports that influence flight fares, such as travel dates, times, and airline carriers. Since, the dataset didn’t have data regarding when the tickets were booked, we will be looking into understanding how airline carriers could influence the price of travel ticket.

Chart

Description automatically generated

1. **Which airline is cheaper?**

Comparing prices across multiple airlines can be an important factor in reducing travel expenses. By comparing prices across multiple airlines, you can find the best deal for your desired route, travel dates, and travel preferences.

The dataset contained 11 different aircraft carriers and the tickets wasn’t evenly distributed across the carriers. The dataset was then sampled using stratified sampling and mean prices were taken to build the bar chart. Mean value was considered instead of median value because the dataset was cleaned, outliers were removed, and data was normalized as well.

On performing this exploratory data analysis, we can conclude that Business class Aircrafts and Premium aircrafts are costlier when compared to economical carriers like SpiceJet, Indigo and GoAir.

Chart, bar chart

Description automatically generated

**Week 7 assignment**

**Purpose:**

To provide a demonstration of your team’s project in this course

**Tasks:**

1. Before making the final submission make sure that you have implemented all my feedbacks from week 1 to week 6.
2. Please ensure that the title page, document control section, table of contents and the reference section is updated and is up to date. Any appendix section should also be updated.
3. Each team member should add an experience statement addressing the following (1) Did this study help you in achieving the objective of this course, (2) Did this study help you in applying all the knowledge you have gained in this course, (3) What were the difficulties (issues) encountered by you during this study and how did you resolve those issues.
4. The team should prepare a 35-40 minutes video demonstration of the study.
5. The demonstration should include a power point presentation.
6. Every team will have an opportunity to go through the demonstration of the predictive analytics system designed by other teams.
7. The team should submit the template document, dataset(s), code/script(s) and the power point presentation deck in the box account set up by the instructor. Details about the box account will be provided in the class or by e-mail.
8. Team members should make all submissions latest by February 23, 2023, at 11:59 PM EST.
9. More instructions (if needed) will be provided by the instructor over e-mail or in class.

Submit your completed week 7 assignment (the updated template document) to the dropbox in Canvas.