**Data Analytics Capstone Topic Approval Form**

**Student Name:** Luke Dorsett

**Student ID:** 011724120

**Capstone Project Name:** Random Forest Regression on Flight Ticket Dataset

**Project Topic**: Multivariate Regression Analysis for Ticket Prices

**☒ This project does not involve human subjects research and is exempt from WGU IRB review.**

**Research Question:** Can a Random Forest Regressor model predict flight ticket sales?

**Hypothesis**:

**Null hypothesis**-H0: A Random Forest Regressor model cannot be made on the research dataset. **Alternate Hypothesis**-Ha: A Random Forest Regressor model can on the selected dataset can reach below 15% MAPE.

**Context:**

The contribution of this study to the field of Data Analytics and the MSDA program is a model predictive model for finding the cost of flights. With the increasing prevalence of flight being used by millions each year, it is often difficult to analyze how expensive flights will be. Individuals needing to fly for work purposes, as well as companies paying for their employees to travel may all encounter difficulties relating to this. This study will create a Random Forest Regressor machine learning model to address this to accurately predict flight prices based on various common factors. A similar study was conducted for this purpose, which experimented with different models on a related dataset. The researchers hypothesized there is promise for a model to be created on this topic, and their research indicated that a Random Forest Regressor is the optimal model to use (Abdella et. al. 2019). In addition, their research indicated that MAPE was the best metric to use. This study will also explore some of the relationships between the independent variables in the dataset and the dependent variable by employing multiple linear regression analysis on the dataset. This will allow for deeper insight into the relationships between variables, so the leading factors behind the price variation will hopefully be revealed.

**Data:**

The research dataset is a collection of flight ticket data from EaseMyTrip.com. The flights are exclusively within India between the dates of February 11th, 2022, to March 31st, 2022, and the dataset in total is 300,261 rows.

The data was collected by the user Shubham Bathwal and is posted on Kaggle.com for public use.

<https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction/data>

|  |  |  |
| --- | --- | --- |
| Serial Number | Continuous | Independent Variable |
| Airline | Categorical | Independent Variable |
| Flight Code | Categorical | Independent Variable |
| Source City | Categorical | Independent Variable |
| Departure Time | Categorical | Independent Variable |
| Stops | Categorical | Independent Variable |
| Arrival Time | Categorical | Independent Variable |
| Destination City | Categorical | Independent Variable |
| Class | Categorical | Independent Variable |
| Duration | Continuous | Independent Variable |
| Days Before | Continuous | Independent Variable |
| Price | Continuous | Dependent Variable |

There are a few advantages to this dataset, along with some disadvantages. One benefit is that the large dataset size is massively impactful for the predictive ability of the model (Couronné et. al. 2018). One limitation is that the data is collected from a small number of airlines. Due to price fluctuations between airlines, the ticket prices may not be generalizable due to revenue disparities. However, this is what makes the data important to analyze, as the algorithm may find these trends, uncovering some of the mystique around air travel (Achyut Joshi, n.d.). Another limitation is that the specific departure and arrival times are not included, only the time of day for both variables. A benefit to possibly offset this is the sheer size of the dataset. A delimitation placed on the study is that additional data will not be researched or scraped and appended to the existing dataset.

**Data Gathering:** The quality of the data is high and has low sparsity at 0%, as seen in its listing on Kaggle.com. The Python tools Pandas and NumPy will be used to convert some of the categorical columns into binary classifications, which will increase the sparsity to approximately 25%, but is needed to run the model. The low data sparsity makes Random Forest Regressors an ideal model as these models work better the lower the sparsity of the data (Xu et. al. 2007). No additional data columns will be added aside from the additional binary columns.

**Data Analytics Tools and Techniques**: To analyze the normality of the data, a Shapiro-Wilk test will be conducted on the dependent variable (price). If the results of the data are not normal, then action will be taken to normalize the data. Afterwards, multiple linear regression will be conducted to analyze the relationships between the independent and dependent variables. Linear regression is optimal for this as it will show detailed information on the statistical significance and impact between the variables used (Alexopoulos 2010). The goal is to then create a Random Forest Regressor model with <15% MAPE, indicating a reasonably accurate model has been achieved which predicts flight tickets with under 15% error. To assist with evaluating the model, a Feature Importance Plot will be created to show the effects of each variable, while Partial Dependence Plots will be used for finding deeper relationships between the selected independent variable and the target variable. The presentation layer will consist of a report describing the findings of the study. Multivariate graphs for linear regression, Partial Dependence Plots, and Feature Importance Plots will be used in the presentation layer.

**Justification of Tools/Techniques:** Python will be used for the duration of this project for numerous reasons. One is the ease of use for creating a professional notebook, which is most easily done with only one language being used at a time. SAS will not be used for good reasons, as there is no built-in MAE or MAPE function available, which is inconvenient for some studies including this one (Brittain et. al. 2018). R is a viable option to perform this study and is extremely powerful when it comes to data analytics on large datasets. However, Python is more widely used within the field of data analytics due to its ease of use and wide range of tools, so it will be used for this study (Colliau et. al. 2017).

**Project Outcomes**: The project will seek to create a Random Forest Regression model for predicting the price of flight tickets in India. Support for the alternative hypothesis is found in Lal et. al. (2023) where the research suggests that a Random Forest model can make the most accurate predictions of other proposed models due to its strength with the type of data being used.

**Projected Project End Date**: 3/30/24

**Sources**:

Abdella, J. A., Zaki, N., Shuaib, K., & Khan, F. (2019). Airline ticket price and demand prediction: A survey. *Journal of King Saud University - Computer and Information Sciences*, *33*(4). <https://doi.org/10.1016/j.jksuci.2019.02.001>

Achyut Joshi. (n.d.). Achyut Joshi. Retrieved March 12, 2023, from <https://achyutjoshi.github.io/btp/flightprices>

Alexopoulos, E. C. (2010). Introduction to Multivariate Regression Analysis. *Hippokratia*, *14*(1), 23–28. Retrieved March 7, 2023, from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049417/>

Brittain, J., Cendon, M., Nizzi, J., & Pleis, J. (2018). Data Scientist’s Analysis Toolbox: Comparison of Python, R, and SAS Performance. *SMU Data Science Review*, *1*(2). Retrieved March 7, 2023, from <https://scholar.smu.edu/cgi/viewcontent.cgi?article=1021&context=datasciencereview>

Colliau, T., Rogers, G., Hughes, Z., Ozgur, C., Hughes, Z., Bennie, E., & Myer-Tyson, quot; (2017). MatLab vs. Python vs. R MatLab vs. Python vs. R. *Journal of Data Science*, *15*, 355–372. Retrieved March 7, 2023, from <https://scholar.valpo.edu/cgi/viewcontent.cgi?article=1049&context=cba_fac_pub>

Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, *19*(1). Retrieved March 7, 2023, from <https://doi.org/10.1186/s12859-018-2264-5>

Dewancker, I., McCourt, M., & Clark, S. (2016, December 14). *Bayesian Optimization for Machine Learning : A Practical Guidebook*. ArXiv.org. Retrieved March 20, 2023, https://doi.org/10.48550/arXiv.1612.04858

Lal, V., Stynes, P., & Cristina Hava Muntean. (2023). An Investigation into Predicting Flight Fares in India Using Machine Learning Models. *Lecture Notes in Networks and Systems*, 106–118. Retrieved March 14, 2023, from <https://doi.org/10.1007/978-3-031-42317-8_9>

Xu, P., & Jelinek, F. (2007). Random forests and the data sparseness problem in language modeling. *Computer Speech & Language*, *21*(1), 105–152. Retrieved March 7, 2023, from <https://doi.org/10.1016/j.csl.2006.01.003>

**Course Instructor Signature/Date:**

☐ The research is exempt from an IRB Review.

☐ An IRB approval is in place (provide proof in appendix B).

Course Instructor’s Approval Status: Approved

Date: Click here to enter a date.

Reviewed by:

Comments: Click here to enter text.