**Data Analytics Capstone Topic Approval Form**

**Student Name:** Luke Dorsett

**Student ID:** 011724120

**Capstone Project Name:** Random Forest Regression for Price of Flight Tickets

**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:** How accurately can a random forest regressor model predict flight ticket sales?

**Hypothesis**:

**Null hypothesis**-H0: A random forest regressor model on the selected dataset cannot reach below 15% MAPE. **Alternate Hypothesis**-Ha: A random forest regressor model on the selected dataset can reach below 15% MAPE.

**Context:**

This study aims to contribute to the data analytics field and the MSDA program by creating a model for predicting 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 they concluded that a random forest regressor is the optimal model to use (Abdella et. al. 2019). 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 it is focuses only on a select few airlines in India, which may not generalize as well to other airlines. Airlines have internal pricing systems which they employ, and this will possibly vary between nations. 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. It is possible that existing data is available on the internet which could show exact departure time, as well as other important variables, but for this study, only the original dataset will be used.

**Data Gathering:** The data will be downloaded from Kaggle.com as a CSV file, including all columns present in the data table without any need for additional transformations. 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, but is needed to run the model. However, the overall sparsity of the data will be relatively low still, making 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). This will be done using Seaborn and Statsmodels.api in Python. In addition, PCA will be used to further evaluate the most important variables to the analysis. Based on the impact of the variable, it may be removed from the analysis to prevent negative effects on the final model. The goal is to then create a random forest regressor model with <15% mean absolute error, indicating a reasonably accurate model. The model will be tested on 20% of the data which will have been set aside for testing, while the other 80% the data will be used for training the model. The presentation layer will consist of a well-documented and thoroughly descriptive notebook file (.ipynb) with visualizations and graphs displayed for an appealing presentation for the viewer.

**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, the ease of use for Python’s Scikit-learn and volume of information of Statsmodels.api for regression cannot be understated, and as a result will be used for the entirety of this study (Colliau et. al. 2017).

**Project Outcomes**: The proposed study aims to create a random forest regressor model for the price of Indian flight tickets with <15%. MAPE is chosen over RMSE or MSE because of the more universal usage of the metric, as it does not depend on the units used, and works well with random forests (Chicco et. al. 2021). This model will contribute to the existing body of work on flight ticket prediction and will hopefully create an accurate model.

**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. <https://achyutjoshi.github.io/btp/flightprices>

Alexopoulos, E. C. (2010). Introduction to Multivariate Regression Analysis. *Hippokratia*, *14*(1), 23–28. <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). <https://scholar.smu.edu/cgi/viewcontent.cgi?article=1021&context=datasciencereview>

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*(5), e623. ncbi. <https://doi.org/10.7717/peerj-cs.623>

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. <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). <https://doi.org/10.1186/s12859-018-2264-5>

Xu, P., & Jelinek, F. (2007). Random forests and the data sparseness problem in language modeling. *Computer Speech & Language*, *21*(1), 105–152. <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.