# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Ferrari Agustin Pablo |
| **Project Name** | Ferrari\_Agustin-24704114-Assessment3 |
| **Date** | 10/11/2023 |
| **Deliverables** | Part 1 - Loading and processing the dataset - Second Iteration.ipynb  Part 2 - Modelling LinearModel and KNN.ipynb |

|  |  |
| --- | --- |
| 1. **EXPERIMENT BACKGROUND** | |
| Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach. | |
| **1.a. Business Objective** | The objective is to build an application that provides users in the USA with quick and accurate price estimates for local flights. Users can enter details about a trip, and the app will immediately predict fares.  This product will assist in travel planning and budgeting, making flight costs clearer for consumers. This service endeavours to empower users to make more informed decisions, potentially leading to cost savings on travel expenses, providing an improved understanding of fare trends across different times and routes.  While the flight price prediction application aims to enhance users' travel planning experience, potential downsides must be considered. Inaccuracy in price estimates may lead to a loss of trust among users, as they rely on the app for budgeting and decision-making. The dynamic nature of airline pricing, influenced by factors beyond the app's control, presents challenges in providing consistently accurate predictions. |
| **1.b. Hypothesis** | **Null Hypothesis (H0):**  These models do not provide more helpful predictions than what is already known or easily determined without them.  **Alternative Hypothesis (H1):**  These models offer more helpful predictions than what is already known or easily determined without them. |
| **1.c. Experiment Objective** | **Project Objective:**  The primary aim of this initiative is to attain a high accuracy rate in predicting local travel airfare. Our focus is on developing models that consistently deliver precise estimates, enhancing users' ability to plan and budget for their trips effectively. The overarching goal is to empower users with reliable information.  **Possible Scenarios Resulting from the Experiment:**  **Good Scenario: High Prediction Accuracy and User Satisfaction:**  In this scenario, the experiment reveals that the predictive models consistently provide accurate estimates for local travel airfare. Users experience a high level of satisfaction with the application, as the predictions align closely with actual prices. This outcome not only enhances users' confidence in the app but also positions it as a reliable tool for travel planning, contributing to positive reviews and increased adoption.  **Normal Scenario:**  In this scenario, the experiment results indicate that the predictive models perform a little better than random. While the models may not significantly outperform simple methods, they still contribute moderately accurate predictions.  **Bad Scenario: Unreliable Predictions and User Distrust:**  In this less favorable scenario, the experiment shows that the predictive models produce inconsistent and inaccurate estimates for local travel airfare. |

|  |  |
| --- | --- |
| 1. **EXPERIMENT DETAILS** | |
| Elaborate on the approach taken for this experiment. List the different steps/techniques used and explain the rationale for choosing them. | |
| **2.a. Data Preparation** | This trial marks the second phase of the Week 1 experiment. Please consult the report for the following aspects:  **Exploratory Data Analysis (EDA): Checkign Duplicate Values**  **EDA: Verification of Data Types.**  **Addressing Inaccurate and Missing Values in String Columns.**  **Identification and Management of Missing Values.**  **Identification of Incorrect or Out-of-Range Values (Numerical).**  **Splitting and cleansing strings. Most of them contained several values divided by || symbols.**  **Standardisation.**  **Division of Data into Training and Testing Sets.** |
| **2.b. Feature Engineering** | Feature engineering is a critical step in refining machine learning models, and six new features have been added to the existing 27 to enhance accuracy. These additions aim to elevate the significance of the data, specifically addressing the need for minimal user input as mandated by business requirements. One of the introduced features, "days\_in\_advance," assesses the time between the flight search date and the actual flight date, considering how airlines adjust pricing based on booking timing. Additionally, "search\_date" and "flight\_date" represent day-of-the-week, month, and year details extracted from flightDate and searchDate, providing a numerical understanding of temporal aspects. "Departure\_time" is derived from segmentsDepartureTimeRaw, focusing on the 24-hour time format for numerical ease. "SegmentsCabinCode" serves as an intermediate step before converting cabin differentiations into percentages, and "numSegments" counts stops in each trip from segmentsCabinCode. Encoding transforms boolean values of "IsRefundable" into integers (1 and 0), making it more useful as numerical input. Lastly, "startingAirport" and "destinationAirport" undergo one-hot encoding to effectively include categorical information. |
| **2.c. Modelling** | In this study, the chosen models for experimentation were the Linear model and K-Nearest Neighbors (KNN). The initial focus of the experiment involved implementing a linear model to quickly model the dataset. Despite its inherent bias, a linear model offers advantages such as easy interpretability, rapid modeling, and a swift method for establishing a baseline. Operating on the assumption that changes in one variable proportionally affect another, linear models are suitable for scenarios with linear relationships between variables. They provide a foundational benchmark, establishing baseline performance metrics for comparison with more intricate models.  GLM was employed as a means of introducing flexibility to the parametric model, especially considering the right-skewness observed in the distribution of the target variable. Both Log of the target variable and inverse gamma distributions were utilised in this context.  Conversely, K-Nearest Neighbors (KNN) is a non-parametric technique employed for both classification and regression. This method makes predictions based on the proximity of data points within the feature space. KNN was chosen for its effectiveness in capturing intricate patterns without necessitating assumptions about the form of the mapping function. The inclusion of KNN complements the linear model, providing a more comprehensive exploration of the dataset and potentially capturing complex relationships that a linear model might overlook. |

|  |  |
| --- | --- |
| 1. **EXPERIMENT RESULTS** | |
| Analyse in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified. | |
| **3.a. Technical Performance** | **Evaluation metric:**  The evaluation metric employed in this project to assess the goodness of fit is the Root Mean Squared Error (RMSE). RMSE is a commonly used metric in regression analysis, measuring the average magnitude of the errors between predicted and observed values. By calculating the square root of the mean squared differences, RMSE provides easier way of understanding the magnitud of the error compared to MSE  **Baseline model:**    **Summary of results:**    The linear models outperformed the baseline, yet they exhibited biased results with a high RMSE, making them less suitable for achieving the business goal.  KNN, on the other hand, demonstrated improved performance, albeit with some indications of overfitting. The optimal Test RMSE was 136.78, a crucial metric for users aiming to precisely predict airfare costs. |
| **3.b. Business Impact** | Business Impact of Failing to Achieve a Positive Outcome, Limited to the Normal Scenario:  In the current scenario, the experimental findings suggest that the predictive models exhibit only marginal improvement over random chance. Although these models might not exhibit a significant advantage over straightforward approaches, they still provide moderately accurate predictions. Despite falling short of achieving a more robust outcome, the models contribute to a certain level of predictive accuracy it may be useful for some users, wanted to roughly estimate a fare. |
| **3.c. Encountered Issues** | The temporal constraint remains a consistent factor, as machine learning models demand a significant amount of time for computations and processing. It's noteworthy that this dataset is exceptionally large, exceeding 13 million rows.  Additionally, not all models could be easily trained. Despite KNN exhibiting better memory handling, the dataset had to be resampled to 20% to facilitate the computation of linear models and prevent memory exhaustion. |

|  |  |
| --- | --- |
| 1. **FUTURE EXPERIMENT** | |
| Reflect on the experiment and highlight the key information/insights you gained from it that are valuable for the overall project objectives from a technical and business perspective. | |
| **4.a. Key Learning** | Key Learning:  The inflexibility of linear models in capturing the underlying function was evident. Conversely, KNN showed superior performance, albeit with signs of overfitting. However, the metrics in both cases fell short of achieving a favorable scenario. |
| **4.b. Suggestions / Recommendations** | 4.b. Suggestions / Recommendations:  We suggest continuing with more experiments. It is advisable to leverage a machine learning model capable of efficiently handling large datasets in terms of computations and memory usage. Light GBM stands out as a suitable candidate for this purpose. Additionally, considering the presence of errors in certain levels of Y, XGBoost, known for its ability to learn from previous mistakes, could be a promising choice.  The errors are easily visualised in the real versus predicted plot for KNN, illustrated for the optimal K parameter model as follows: |