# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Divgun Singh |
| **Project Name** | Assignment 3 |
| **Date** | 10/11/2023 |
| **GitHub Repo** | https://github.com/divgunsingh/Airline\_price\_modelling.git |
| **Deliverables** | * Keras\_DNN\_DivgunSingh\_24586556.ipyng * Handling Missing values * One Hot Encoding of categorical features * New Features based on the available features to enhance training * Keras Sequential Deep Neural Networks Model * RMSE and MAE evaluation metrics |

|  |  |
| --- | --- |
| 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** | Explain clearly what is the goal of this project for the business. How will the results be used? What will be the impact of accurate or incorrect results?   * The main goal of the experiment is to accurately predict the prices of airline travel based on the inputs captured by the user like Starting Airport, Destination Airport, Departure Date, Cabin Type etc. * This will be helpful not only for the airline business to reduce the load on customer service calls for pricing and availabilities but also will be relaxing for the consumers directly helping them save their worthy time. |
| **1.b. Hypothesis** | Present the hypothesis you want to test, the question you want to answer or the insight you are seeking. Explain the reasons why you think it is worthwhile considering it,   * In this project, I will be checking the total airline price after capturing various user inputs and based on the inputs I will try to create new features based on the available data from the airline industry to predict the resulting price. * Also, I will be checking for specific days of month or months of year when the demand is really high so the price also increases because of limited supply and hence people end up paying higher prices for the same travel. * It is worthwhile considering this hypothesis so that data can be understood and analyzed in a way that user can have an estimate of money he/she will be spending for their future travel and can plan their trip dates accordingly and also compare the prices on different dates. |
| **1.c. Experiment Objective** | Detail what will be the expected outcome of the experiment. If possible, estimate the goal you are expecting. List the possible scenarios resulting from this experiment.   * The expected outcome of the model will be an application where the users can get the prices of their air fare based on the inputs they will be providing in the application * Based on the inputs entered, the model will predict total airfare for their travel on a particular date between their starting and destination airports with their specified cabin types. * The results will help businesses to put less load and hire less resources for their on call customer service and also to the users to have an estimate idea of their trip while planning as they will know their travel costs beforehand. |

|  |  |
| --- | --- |
| 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** | Describe the steps taken for preparing the data (if any). Explain the rationale why you had to perform these steps. List also the steps you decided to not execute and the reasoning behind it. Highlight any step that may potentially be important for future experiments   1. Data Cleaning: Checked for missing values in features, identified nulls in totalTravelDistance and segmentsEquipmentDescription, and omitted those columns. Detected and omitted further null values in segmentsDistance. 2. Processing "segment" Columns: Assessed and handled 12 features containing segmented flight data separated by "II" individually. 3. Encoded segmentCabinCode into individual features, assigning percentages based on their presence in a specific row. 4. Encoding Date Variables: Encoded departureDate and arrivalDate into DepartureDateDayOfWeek and DepartureDateMonth. 5. Processed segmentsDepartureTimeRaw to extract the first datetime, computed the day of the week and month for each timestamp. 6. Encoding isRefundable: 7. Transformed isRefundable into numerical format by mapping Boolean values (True and False) to integers (1 and 0). 8. Omitted Features: Excluded features like lagId, isBasicEconomy, isNonStop, totalTravelDistance, travelDuration, and others due to redundancy or impracticality. 9. Scaling: Scaled all values using StandardScaler to prevent features from being weighed differently based on their ranges. 10. Feature Engineering (Six Supplementary Features):     1. days\_in\_advance: Examines the number of days between the date of the flight's search and the actual flight date.     2. search\_date: Extracted day\_of\_the\_week, month, and year from flightDate, converted values into an integer format.     3. flight\_date: Extracted day\_of\_the\_week, month, and year from searchDate, converted values into an integer format.     4. departure\_time: Extracted from segmentsDepartureTimeRaw, originally formatted as ('2022-05-20T18:58:00.000-07:00||2022-05-21T00:5...').     5. segmentsCabinCode: Extracted from segmentedCabinCode, acts as an intermediate feature prior to converting cabin differentiations into percentages.     6. numSegments: Count of stops in each trip, derived from segmentsCabinCode. 11. startingAirport: One-hot encoding applied. 12. destinationAirport: One-hot encoding performed. |
| **2.b. Feature Engineering** | Describe the steps taken for generating features (if any). Explain the rationale why you had to perform these steps. List also the feature you decided to remove and the reasoning behind it. Highlight any feature that may potentially be important for future experiments   * The steps taken to create new features included extracting day, month, year and date from the dates present in the dataset. Also, the number of days for which the flight was booked in advance was calculated based on the flight date and the current date when it is getting booked. * In the future it will be important if more features can be created based on the different segments of arrival and departure given to enhance the model and bring more accurate results. |
| **2.c. Modelling** | Describe the model(s) trained for this experiment and why you choose them. List the hyperparameter tuned and the values tested and also the rationale why you choose them. List also the models you decided to not train and the reasoning behind it. Highlight any model or hyperparameter that may potentially be important for future experiments   * A Keras Sequential Deep Neural Network (DNN) model was chosen for airline price prediction due to its ability to capture intricate patterns and non-linear relationships in the complex dataset. The model's feature learning capabilities, flexibility in architecture design, and suitability for regression tasks align well with the diverse input features influencing airline prices. With the availability of historical pricing data, the DNN is poised to continuously improve predictions and adapt to evolving patterns, making it a robust choice for accurate and dynamic airline price forecasting * The Dense layers define the neural network's structure. The first Dense layer with 64 units and ReLU activation allows for learning complex patterns. The second Dense layer with 32 units may capture less complex features. The output Dense layer with 1 unit and no activation is appropriate for regression tasks. The choice of hyperparameters, such as the number of units, activation functions, and layer types, balances model complexity and computational efficiency to achieve accurate predictions for the airline price regression task.. |

|  |  |
| --- | --- |
| 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** | Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes.   * The model was assessed and analyzed using the metric RMSE which stands for Root Mean Square Error, and it is a common metric used to evaluate the accuracy of a predictive model, particularly in prediction problems. RMSE measures the average magnitude of the errors or the differences between predicted values and actual values in a dataset.  |  |  |  |  |  | | --- | --- | --- | --- | --- | | **Model** | **Validation Set (MAE)** | **Validation Set (RMSE)** | **Test Set (MAE)** | **Test Set (RMSE)** | | DNN (Keras Sequential) | 81.52799 | 121.19491 | 81.52467 | 121.08997 | |
| **3.b. Business Impact** | Interpret the results of the experiments related to the business objective set earlier. Estimate the impacts of the incorrect results for the business (some results may have more impact compared to others)   * As far as the business is concerned with the wrong forecasting of total fare, users will be giving bad reviews to the airline company and also will be troubling the business during the checkout phase of the flight booking. This will create a huge load for giving answers by the on-call customer support team and business will take a huge loss because of bad market reputation which might result in the plunge of the stock prices of the airline company too. |
| **3.c. Encountered Issues** | List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them. Highlight also the issues that may have to be dealt with in future experiments.   * The major issue faced was related to the compute and memory resources to deal with such large amounts of dataset. Because of this additional compute resources were purchased on colab pro. |

|  |  |
| --- | --- |
| 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** | Reflect on the outcome of the experiment and list the new insights you gained from it. Provide rationale for pursuing more experimentation with the current approach or call out if you think it is a dead end.   * I will try to make more advances features based on the features already present in the dataset to find out more quality trained dataset and also fine-tune the models with more advances hyperparameter tuning. * Also, will explore more advanced prediction models among the deep neural networks and will also explore the power of random forest for this problem. |
| **4.b. Suggestions / Recommendations** | Given the results achieved and the overall objective of the project, list the potential next steps and experiments. For each of them assess the expected uplift or gains and rank them accordingly. If the experiment achieved the required outcome for the business, recommend the steps to deploy this solution into production. |