# EXPERIMENT REPORT

| | |
|---|---|
| **Student Name** | Smit Dipeshkumar Khatri |
| **Project Name** | AT3 - Data Product with Machine Learning |
| **Date** | 10/11/2023 |
| **Deliverables** | <Jupyter><br><Decision Tree Regressor> |

---

| **1. EXPERIMENT BACKGROUND** | |
|---|---|
| | |
| **1.a. Business Objective** | The primary business objective is to develop an accurate flight fare prediction model to empower travelers and enhance revenue management for airlines. This model aims to provide travelers with reliable fare estimates, enabling them to make informed decisions when booking flights. Simultaneously, it assists airlines in optimizing pricing strategies based on demand forecasts and market dynamics. Ultimately, the goal is to improve the overall travel experience for consumers while increasing the profitability and efficiency of airlines. |
| **1.b. Hypothesis** | Hypothesize that by implementing a Decision Tree Regressor model with one-hot encoding for data preprocessing, we can accurately predict flight fares. This model will capture complex relationships in the data, allowing us to provide travelers with precise fare estimates. We anticipate that the model's ability to generalize to unseen data will lead to improved decision-making for travelers and enhanced revenue management for airlines. |
| **1.c. Experiment Objective** | The primary objective of this experiment is to develop and evaluate a machine learning model capable of accurately predicting flight fares. The model will be based on a Decision Tree Regressor and will incorporate data preprocessing techniques such as one-hot encoding. By successfully achieving this objective, we aim to provide a valuable tool for travelers to estimate flight costs effectively. Additionally, we intend to explore the potential of the model to support airlines in optimizing their pricing and revenue management strategies. The experiment's success will be measured by assessing the model's performance through relevant regression metrics. |

| 2. EXPERIMENT DETAILS | |
|---|---|
| | |
| **2.a. Data Preparation** | The data preparation phase involved several crucial steps to ensure the quality and suitability of the dataset for machine learning. These steps included data cleaning, feature selection, and encoding categorical variables.<br><br>Data Cleaning: The initial dataset was thoroughly examined to identify and handle missing values, outliers, and inconsistencies. Cleaning the data was essential to prevent any noise or bias from affecting the model's performance.<br><br>Feature Selection: We carefully selected relevant features for our model, focusing on those with a direct influence on flight fares. Features like the starting and destination airports, flight date, departure time, and cabin code were retained for modeling.<br><br>Categorical Variable Encoding: Categorical variables, such as airport codes and cabin codes, were one-hot encoded to transform them into a numerical format suitable for machine learning. This encoding step enabled the model to process these variables effectively. |
| **2.b. Feature Engineering** | Feature engineering is a critical aspect of building an effective machine learning model for flight fare prediction. The process involved creating new features and transforming existing ones to enhance the model's ability to capture underlying patterns and relationships in the data.<br><br>Date-Based Features: To leverage the temporal aspect of flight fares, we derived additional features from the 'flightDate.' These included 'flightYear,' 'flightMonth,' 'flightDay,' and 'flightWeek.' These features can help the model understand how fares vary with time.<br><br>Duration and Distance Features: We computed 'travelDuration' and 'totalTravelDistance' features to capture the flight's length and distance. Longer flights tend to have higher fares, and these features can help the model account for such variations.<br><br>Time-of-Day Features: The 'segmentsDepartureTimeRaw' feature was transformed to extract the time of day when the flight departs. This can be a crucial factor in fare determination, as flights during peak hours may cost more.<br><br>Categorical Variable Encoding: Categorical variables, such as 'startingAirport,' 'destinationAirport,' and 'segmentsCabinCode,' were one-hot encoded to represent them as binary values. This transformation allows the model to consider different categories for these variables.<br><br>Interaction Features: We created interaction features to capture potential relationships between different input features. For instance, the interaction between the 'startingAirport' and 'destinationAirport' could reveal insights into specific route fares. |

| | |
|---|---|
| **2.c. Modelling** | The core of our flight fare prediction project involves the development of a predictive model to estimate airline fares. The chosen machine learning algorithm is the Decision Tree Regressor, which is implemented using the scikit-learn library. Here's a brief overview of the modelling process:<br><br>Algorithm Selection: Decision trees are versatile models that can capture both linear and non-linear relationships in the data. They are particularly suitable for regression tasks, making them an excellent choice for predicting flight fares. The scikit-learn library offers a robust implementation of decision tree regressors.<br><br>Data Preprocessing: Before feeding the data into the model, we must prepare it. This includes handling categorical variables using one-hot encoding through a ColumnTransformer. This transformation allows the model to work with binary representations of categorical variables.<br><br>Data Split: To evaluate the model's performance, we divided our dataset into two parts: a training set (80% of the data) and a testing set (20% of the data). The training set is used to train the model, while the testing set assesses how well the model generalizes to unseen data. This random split minimizes the risk of overfitting.<br><br>Model Development: The Decision Tree Regressor is integrated into a scikit-learn pipeline. This pipeline combines the data preprocessing and the model into a single entity, streamlining the process. The training set is used to fit the model to the data, allowing it to learn patterns and relationships.<br><br>Evaluation Metrics: To assess the model's performance, we utilize two common regression metrics: Mean Squared Error (MSE) and Mean Absolute Error (MAE). MSE quantifies the average squared difference between predicted and actual fares, while MAE measures the average absolute difference. The lower the values of these metrics, the better the model's predictive accuracy.<br><br>Model Evaluation: The model's performance is evaluated based on the MSE and MAE values, providing insights into how well it predicts airline fares. In our case, the model achieved a Root Mean Squared Fare Prediction of $138.97 and a Mean Absolute Fare Prediction of $97.54.<br><br>Model Deployment: Once the model is developed and evaluated, it can be deployed for various applications. This includes integrating it into online booking systems or fare comparison websites, helping users estimate flight costs more accurately. |

| 3. EXPERIMENT RESULTS | |
|---|---|
| **3.a. Technical Performance** | The technical performance of our flight fare prediction model, based on a Decision Tree Regressor, reflects its effectiveness in accurately estimating airline fares. Here's an overview of its technical performance:<br><br>Model Accuracy: The primary measure of our model's technical performance is its accuracy in predicting flight fares. This is determined through metrics such as Mean Squared Error (MSE) and Mean Absolute Error (MAE). Our model achieved a Root Mean Squared Fare Prediction of $138.97 and a Mean Absolute Fare Prediction of $97.54. These values indicate how well the model aligns with actual fare data, with lower values representing better accuracy.<br><br>Generalization: Our model demonstrates good generalization to unseen data. This is confirmed through the separation of our dataset into a training set (used for model training) and a testing set (used for evaluation). The model's performance on the testing set reflects its ability to make accurate predictions beyond the data it was trained on.<br><br>Scalability: The model's scalability depends on the underlying decision tree algorithm and its implementation. Decision tree models can be efficiently trained on large datasets, making them suitable for handling a growing volume of airline fare data.<br><br>Speed and Efficiency: Decision tree models are relatively fast for prediction tasks, which is crucial for applications requiring real-time fare estimation, such as online booking systems.<br><br>Interpretability: Decision tree models are highly interpretable, making it easier to understand the factors influencing fare predictions. This transparency is essential for both technical and non-technical stakeholders.<br><br>Data Handling: The model is equipped to handle different types of data, including categorical features that have been one-hot encoded. It accommodates a wide range of input features, enabling robust fare predictions. |
| **3.b. Business Impact** | The implementation of an accurate flight fare prediction model holds significant potential for various stakeholders in the travel industry. Here's a brief overview of the business impact of our model:<br><br>Enhanced User Experience: By providing travelers with more accurate fare estimates, airlines and travel agencies can improve the user experience. Travelers can make more informed decisions and plan their trips with greater confidence.<br><br>Increased Customer Loyalty: Improved fare prediction leads to higher customer satisfaction, which can translate into increased loyalty. Travelers are more likely to return to airlines or booking platforms that consistently provide reliable fare estimates.<br><br>Competitive Advantage: Airlines and travel agencies that deploy our fare prediction model gain a competitive edge in the market. Accurate pricing helps attract more customers and boosts revenue.<br><br>Dynamic Pricing Optimization: The model can be used to optimize dynamic pricing strategies. Airlines can adjust fares based on demand, time of booking, and other factors, maximizing profitability.<br><br>Reduced Revenue Loss: Accurate fare prediction minimizes instances of customers |

| | |
|---|---|
| | abandoning bookings due to unexpected price fluctuations. This reduction in revenue loss directly impacts the bottom line.<br><br>Operational Efficiency: Streamlined fare prediction contributes to operational efficiency. Airlines can better manage capacity and allocate resources effectively.<br><br>Marketing Insights: The model's interpretability provides valuable insights into the factors affecting fare changes. This information can guide marketing and sales strategies.<br><br>Customer Retention: The ability to offer fair and consistent pricing encourages customer retention. Satisfied customers are more likely to choose the same airline or booking platform for future travel needs.<br><br>Data-Driven Decision Making: Airlines and travel agencies can make data-driven decisions regarding pricing, promotions, and route planning. This improves business strategies and profitability. |
| **3.c. Encountered Issues** | During the development and implementation of the flight fare prediction model, several challenges and issues were encountered. These issues included:<br><br>Data Quality and Consistency: The dataset used for training and testing the model contained inconsistencies and missing values. This required data preprocessing efforts, such as data cleaning and imputation, to ensure the quality and reliability of the model.<br><br>Feature Engineering Complexity: Selecting the most relevant features and engineering them effectively was a complex task. Deciding which variables have the most significant impact on fare prediction required in-depth analysis and domain knowledge.<br><br>Model Generalization: Ensuring that the model could generalize to unseen data was a critical concern. Careful selection of training and testing datasets, as well as model validation techniques, was necessary to address this issue.<br><br>Overfitting: Decision tree models are prone to overfitting. Managing the model's complexity and preventing it from fitting noise in the data were important to ensure accurate fare predictions.<br><br>Data Volume: The volume of available historical flight data was limited. More extensive datasets could potentially lead to even more accurate predictions.<br><br>Model Interpretability: Decision trees, while interpretable, can become complex. Simplifying the model for better interpretability without sacrificing predictive power was a constant consideration.<br><br>External Factors: External factors such as economic conditions, geopolitical events, and natural disasters can impact flight fares. These external variables were not included in the dataset but may play a significant role in fare prediction.<br><br>Deployment and Integration: Transitioning the model from a development environment to a production environment and integrating it with existing booking systems or platforms required careful planning and implementation.<br><br>Hyperparameter Tuning: Optimizing the model's hyperparameters to achieve the best performance was an iterative process. Finding the right balance between model complexity and predictive accuracy was challenging. |

| | |
|---|---|
| | Scalability: As the volume of flight bookings increases, the model must be able to scale to handle a larger workload efficiently. |

| 4. FUTURE EXPERIMENT | |
|---|---|
| | |
| **4.a. Key Learning** | The development and implementation of the flight fare prediction model have yielded several valuable insights and learnings. These key takeaways include:<br><br>Data Quality is Paramount: High-quality data is fundamental to building accurate predictive models. Cleaning, preprocessing, and ensuring data consistency are critical steps in the data preparation process.<br><br>Feature Engineering is a Delicate Art: Feature selection and engineering require a deep understanding of the domain and the impact of variables on the target. Well-engineered features can significantly enhance model performance.<br><br>Model Selection Matters: The choice of the machine learning algorithm is crucial. Decision Tree Regressors are powerful and interpretable, but other algorithms should be considered, especially for complex datasets.<br><br>Data Splitting Strategy: The way data is split into training and testing sets has a substantial impact on model evaluation. Random splits with cross-validation can help validate a model's ability to generalize.<br><br>Model Interpretability: While complex models can provide high predictive accuracy, interpretability is equally important. Decision trees offer transparency and are valuable for understanding how predictions are made.<br><br>External Factors Play a Role: Factors external to the dataset, such as economic conditions and geopolitical events, can influence fare prices. Consideration of these variables can enhance prediction accuracy.<br><br>Hyperparameter Tuning is Iterative: Optimizing a model's hyperparameters is often an iterative process. Experimenting with different settings and evaluating their impact on performance is necessary.<br><br>Deployment Challenges: Transitioning from a development environment to a production environment is a complex task. Integration with existing systems and ensuring scalability are important considerations.<br><br>Potential for Business Impact: Accurate fare predictions can positively impact the travel industry by providing travelers with more reliable cost estimates and helping airlines optimize pricing strategies.<br><br>Continuous Improvement: Model development is an ongoing process. Regularly |

| | |
|---|---|
| | updating and retraining the model with new data can help maintain and improve its predictive accuracy. |
| **4.b. Suggestions / Recommendations** | Based on the insights gained from this flight fare prediction project, here are some recommendations and suggestions for further enhancing the model and its potential impact:<br><br>Incorporate External Data: Consider integrating external data sources that may influence flight fares, such as fuel prices, economic indicators, and seasonal trends. This can lead to more accurate predictions.<br><br>Advanced Algorithms: Explore more advanced regression algorithms, such as Random Forest, Gradient Boosting, or Neural Networks. These models can capture complex relationships within the data and may lead to improved performance.<br><br>Hyperparameter Tuning: Invest more time in hyperparameter tuning to optimize the selected algorithm. Techniques like grid search or random search can help identify the best parameter combinations.<br><br>Feature Engineering: Continue to refine and expand feature engineering efforts. Create new features or derive insights from existing ones to better capture the nuances of flight fare dynamics.<br><br>Interpretability Tools: Implement model interpretability tools to provide insights into how the model makes predictions. This can enhance transparency and build trust with stakeholders.<br><br>Real-Time Pricing: Consider developing a real-time pricing engine that integrates the model, allowing airlines to dynamically adjust fares based on demand and market conditions.<br><br>Deployment Strategy: Plan a robust deployment strategy that addresses scalability, monitoring, and integration with airline reservation systems. Collaborate closely with IT and DevOps teams to ensure a smooth transition to a production environment.<br><br>Regular Model Updates: Establish a process for regularly updating and retraining the model. Flight fare dynamics change over time, and keeping the model current is essential for continued accuracy.<br><br>User-Friendly Interfaces: Create user-friendly interfaces for travelers and travel booking platforms. Providing fare estimates and explanations can enhance the user experience.<br><br>Ethical Considerations: Be mindful of ethical considerations when using predictive models for pricing. Ensure that pricing strategies are fair and non-discriminatory.<br><br>Feedback Mechanism: Implement a feedback mechanism that collects user feedback on predicted fares. This data can be valuable for further model improvement.<br><br>Cross-Functional Collaboration: Foster collaboration between data scientists, domain experts, and business stakeholders. This multidisciplinary approach ensures that the model aligns with business objectives. |

| | |
|---|---|
| | Business Continuity: Develop a robust business continuity plan to address potential disruptions, such as unforeseen external events, that may impact travel and pricing.<br><br>Regulatory Compliance: Stay informed about regulations related to fare pricing and ensure that the model complies with legal and industry standards.<br><br>Educational Initiatives: Educate users and stakeholders about how the model works and what factors influence fare predictions. Transparency can build trust and acceptance.<br><br>Competitive Analysis: Continuously monitor and analyze the fare prediction models of competitors in the travel industry. This can help maintain a competitive edge.<br><br>Measure Impact: Establish key performance indicators (KPIs) to measure the impact of the model on airline revenue, user satisfaction, and pricing strategy efficiency.<br><br>Research and Development: Allocate resources to ongoing research and development to stay at the forefront of fare prediction technology. |