# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Aibarna Singh Basnet |
| **Project Name** | **AT3 - Data Product with Machine Learning** |
| **Date** | 9/11/2023 |
| **Deliverables** | basnet\_aibarna-24585717-randomforest.ipynb  Aibarna\_model.pkl |

|  |  |
| --- | --- |
| 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?  Business Goal:  The project aims to improve user experience for USA-based air travelers by providing accurate airfare predictions through a Streamlit app. This contributes to user satisfaction, retention, and a competitive advantage in the travel planning market.  Impact of Accurate Results:  Accurate predictions build trust, leading to higher user satisfaction, positive brand perception, and increased user engagement. Users are more likely to rely on the app for future travel planning, enhancing overall business success.  Impact of Incorrect Results:  Inaccurate predictions result in a loss of user trust, negative brand image, and reduced user retention. Consistency in providing accurate information is crucial for the app's long-term success and market position. |
| **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,  Hypothesis:  The inclusion of additional contextual factors such as weather conditions and major events (festivals, conferences) in the predictive model will improve the accuracy of airfare estimations in the Streamlit app.  Reasoning:  1. Seasonal Variations: Weather conditions can impact travel demand. For instance, flights during holiday seasons or adverse weather conditions might experience price fluctuations.  2. Event Impact: Major events at the destination can affect airfare prices. An inclusion of event data may capture spikes in demand, leading to more precise predictions.  3. User Preferences: Understanding user preferences during specific seasons or events can enhance personalization, contributing to a more tailored and accurate prediction.  4. Competitive Edge: Incorporating these factors may provide a unique selling point, differentiating the app from competitors and attracting users seeking more comprehensive travel predictions.  By testing this hypothesis, we aim to refine the predictive model, ensuring it considers a broader range of influential factors for more accurate and nuanced airfare estimations. |
| **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.  Expected Outcome:  We aim for a 10% improvement in prediction accuracy by incorporating additional factors such as weather conditions and major events into the predictive model for airfare estimations. This enhancement is anticipated to provide users with more precise predictions, fostering increased trust and satisfaction.  Possible Scenarios:  Success is achieving the estimated goal, while a moderate success scenario involves a noticeable but slightly below-estimated improvement. A neutral scenario sees minimal change, and a challenge scenario indicates a decrease in accuracy. An unexpected discovery scenario would involve identifying previously overlooked influential factors, potentially reshaping the approach to airfare prediction. These scenarios guide our understanding and refinement of the predictive model. |

|  |  |
| --- | --- |
| 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  The data preparation involved several critical steps to enhance its quality and suitability for analysis. Firstly, columns containing '||' delimiters were exploded to increase granularity and simplify subsequent analyses, providing a more detailed perspective on the dataset. Subsequently, relevant date and time features were extracted from the 'segmentsDepartureTimeRaw' column to enable time-based analysis and eliminate redundant information.  To address missing numerical values, imputation was performed based on the mode values specific to airport pairs, ensuring completeness while maintaining context relevance. Categorical features, including 'startingAirport', 'destinationAirport', and 'segmentsCabinCode', were mapped to numerical codes for compatibility with machine learning models.  Downcasting was employed on numerical columns to optimize memory usage, crucial for large datasets where efficiency is paramount. Notably, NaN values in 'segmentsDistance' were not handled separately, as they were addressed during the numerical imputation step.  For potential future experiments, considerations include further exploration of datetime handling, feature scaling for numerical columns, additional feature engineering, cross-validation implementation, and feature importance analysis. These steps could contribute to a more refined understanding of the dataset and improved model performance. |
| **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  Feature generation involved extracting relevant information and transforming the dataset to enhance its suitability for analysis. Key steps included exploding columns with '||' delimiters for increased granularity and extracting date and time features from the 'segmentsDepartureTimeRaw' column, providing a more detailed temporal perspective. Numerical imputation and categorical mapping were employed to handle missing values and convert categorical features into a format suitable for machine learning models.  A feature that was removed during preprocessing was 'totalFare,' as it was replaced by the newly created target variable 'Total\_fare\_mode,' calculated as the mode of total fares based on specific features. This transformation aimed to provide a more robust and representative target for regression modeling.  For future experiments, features such as datetime components ('year', 'month', 'day', 'hour', 'minute') could be crucial for temporal analysis and model performance. Exploring additional feature engineering techniques and assessing feature importance may uncover valuable insights and further enhance the dataset for future experiments. |
| **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  The model trained for this experiment is a Random Forest Regressor, chosen for its ability to handle non-linear relationships and capture complex patterns in the data. The following hyperparameters were tuned with specific values based on considerations for model performance:  1. Number of Estimators (n\_estimators): Set to 50 to balance computational efficiency and model robustness.  2. Maximum Depth (max\_depth): Limited to 10 to prevent overfitting and ensure the model generalizes well.  3. Maximum Features (max\_features): Set to 'sqrt' to consider the square root of the number of features, an effective choice for regression tasks.  4. Number of Jobs (n\_jobs): Utilized all available processors (-1) for parallelized training, improving computational efficiency.  The Random Forest model was selected for its ensemble learning nature, combining multiple decision trees to enhance predictive accuracy and mitigate overfitting. These hyperparameters strike a balance between model complexity and computational efficiency.  No other models were trained for this experiment as Random Forests are well-suited for regression tasks, particularly when dealing with a moderate-sized dataset and diverse feature types. Future experiments may explore alternative ensemble methods or more complex models, such as gradient boosting, depending on the dataset's characteristics and size. Additionally, further hyperparameter tuning and experimentation with different ensemble techniques could be explored for improved model performance in future iterations. |

|  |  |
| --- | --- |
| 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 Root Mean Square Error (RMSE) of the Random Forest Regressor model on the test set is 183.94. This metric quantifies the average magnitude of the errors between predicted and actual values, with lower values indicating better model performance.  Analyzing the underperforming cases, instances where the model's predictions significantly deviate from the actual values may be attributed to several factors:  1. Outliers: Presence of outliers in the dataset can disproportionately influence the RMSE, leading to higher error values. Identifying and handling outliers could improve model robustness.  2. Non-linearity: If the underlying relationship between features and the target variable is highly non-linear, a Random Forest model with limited depth may struggle to capture intricate patterns, contributing to higher errors.  3. Feature Importance: Reassessing feature importance could uncover overlooked variables crucial for accurate predictions. Fine-tuning the feature set based on importance may enhance the model's predictive power.  4. Hyperparameter Tuning: Further optimization of hyperparameters, such as adjusting the number of estimators, maximum depth, or exploring different ensemble techniques, might lead to improved model performance.  5. Data Quality: Ensuring data quality, handling missing values more effectively, and refining preprocessing steps could contribute to a more accurate and reliable model.  Iterative experimentation, including refining features, adjusting hyperparameters, and addressing potential outliers, is essential for enhancing model performance in future iterations. |
| **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)  The current Random Forest Regressor model, with an RMSE of 183.94, provides a moderate level of accuracy for predicting flight fares. However, potential impacts on the business include financial implications, customer dissatisfaction, and a competitive disadvantage. Inaccurate fare predictions may affect operational efficiency, decision support, and harm the airline's reputation. To address these impacts, iterative improvements in the model, focusing on features, outliers, hyperparameters, and data quality, are crucial for aligning predictions with the business objective and minimizing adverse consequences. Regular model updates and monitoring are essential for sustaining accurate fare estimations and mitigating potential business risks. |
| **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.  Issues Faced:  1. Handling NaN Values in 'segmentsDistance':  - Status: Solved during numerical imputation based on airport pairs.  - Solution: Filled NaN values using the mode of 'segmentsDistance' specific to airport pairs.  2. Model Performance (RMSE):  - Status: Unsolved; RMSE indicates moderate accuracy.  - Solution: Iterative experimentation required for hyperparameter tuning, feature refinement, and addressing outliers. Exploring alternative models or ensemble techniques may be considered.  3. Feature Importance:  - Status: Unsolved; potential impact on model accuracy.  - Solution: Reassess feature importance and consider refining the feature set based on its impact on predictive performance.  4. Outliers:  - Status: Identified as a potential issue.  - Solution: Detection and handling of outliers during preprocessing. Investigate more robust outlier detection methods for improved model robustness.  5. Data Quality:  - Status: Partially addressed during preprocessing.  - Solution: Ongoing efforts to enhance data quality, particularly in handling missing values and ensuring consistency in datetime formats.  Issues for Future Experiments:  1. Datetime Handling Consistency:  - Solution: Ensure consistency in datetime formats; explore additional datetime features for improved model understanding.  2. Feature Scaling:  - Solution:Consider scaling numerical features to address potential discrepancies in variable magnitudes.  3. Feature Engineering:  - Solution:Explore additional feature engineering techniques to capture complex relationships in the data.  4. Hyperparameter Tuning:  - Solution:Experiment with different hyperparameter configurations for the Random Forest model to optimize performance.  5. Cross-Validation:  - Solution:Implement cross-validation for a more robust model evaluation, especially if the dataset size allows.  6. Feature Importance Analysis:  - SolutionConduct feature importance analysis to understand the impact of each feature on the model's predictions.  7. Model Selection:  - Solution:Explore alternative models or ensemble techniques based on the dataset's characteristics and size for potential improvements in predictive performance.  Addressing these issues in future experiments involves a combination of model refinement, feature engineering, and thorough analysis of data quality and consistency. Continuous monitoring and adjustment are essential for sustaining and improving model performance. |

|  |  |
| --- | --- |
| 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.  Reflection on Experiment Outcome:  The experiment yielded several insights into predicting flight fares using a Random Forest Regressor. The RMSE of 183.94 indicates moderate accuracy, but it also highlights areas for improvement. Key insights include:  1. Moderate Accuracy: The model provides reasonable predictions, but there is room for enhancement, especially considering the sensitivity of the airline industry to accurate fare estimates.  2. Feature Importance: Reassessing feature importance revealed variables crucial for predictions. This insight guides future feature engineering and selection to improve the model's understanding of relevant factors.  3. Data Quality Challenges Ongoing challenges in data quality, particularly in handling missing values and ensuring datetime consistency, underscore the importance of robust preprocessing for model reliability.  Rationale for More Experimentation:  1. Optimization Potential: The moderate accuracy suggests optimization opportunities through hyperparameter tuning, feature refinement, and outlier handling. Iterative experimentation can uncover the optimal model configuration.  2. Feature Exploration: Continued exploration of feature importance and potential interactions between variables may unveil hidden patterns, contributing to improved predictive performance.  3. Business Value: Accurate fare predictions directly impact business outcomes, affecting revenue and customer satisfaction. The pursuit of more experimentation aligns with the business goal of providing precise fare estimates.  4. Flexibility of Approach: The current approach allows for adjustments in model configuration and feature engineering. This flexibility supports further experimentation with different techniques and models. |
| **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.  Potential Next Steps and Experiments:  1. Hyperparameter Tuning:  - Expected Uplift: High  - Rationale: Fine-tuning hyperparameters such as the number of estimators, maximum depth, and minimum samples per leaf can significantly enhance model performance.  2. Feature Engineering:  - Expected Uplift: Medium to High  - Rationale: Further exploration of feature interactions, creation of new relevant features, and transformation techniques may uncover patterns that improve predictive accuracy.  3. Ensemble Techniques:  - Expected Uplift: Medium  - Rationale: Experimenting with alternative ensemble methods like Gradient Boosting or XGBoost may provide improved predictive power compared to the current Random Forest model.  4. Outlier Detection and Handling:  - Expected Uplift: Medium  - Rationale: Identifying and appropriately handling outliers in the dataset can mitigate their impact on model accuracy, potentially leading to more robust predictions.  5. Cross-Validation:  - Expected Uplift: Medium  - Rationale: Implementing cross-validation during model evaluation can provide a more robust estimate of the model's performance and generalizability.  6. Advanced Time Series Modeling:  - Expected Uplift: High  - Rationale: If temporal patterns play a significant role in fare predictions, exploring advanced time series models such as SARIMA or Prophet may capture these dynamics more effectively.  Deployment Recommendations:  1. Model Evaluation:  - Conduct thorough evaluation of the tuned model on a separate validation dataset to ensure generalizability.  2. Documentation:  - Document the final model configuration, hyperparameters, and preprocessing steps for transparency and reproducibility.  3. Scalability Considerations:  - Assess the scalability of the model to handle larger datasets and increased prediction demands.  4. Integration with Production Systems:  - Integrate the model into the existing production environment, ensuring seamless interaction with other systems.  5. Monitoring and Maintenance:  - Implement a robust monitoring system to track model performance in real-time and schedule periodic model retraining to account for changing patterns.  6. Feedback Loop:  - Establish a feedback loop for continuous improvement, incorporating user feedback and adapting the model as needed. |