**Final Report: Predicting Hotel Reservation Cancellations**

**1. Introduction**

**1.1 Objective**

**The objective of this project is to build a machine learning model to predict whether a hotel reservation will be cancelled based on the features provided. This prediction can help hotels manage their bookings efficiently and reduce revenue losses due to cancellations.**

**1.2 Dataset Overview**

**The dataset contains information about hotel reservations, including the number of adults, children, weekend nights, weekday nights, room type, lead time, market segment, and whether the reservation was cancelled or not.**

**2. Exploratory Data Analysis (EDA)**

**2.1 Data Distribution and Feature Exploration**

* **The dataset was analyzed for null values, duplicates, and class distribution.**
* **Key categorical features were encoded using one-hot encoding, and numerical features were scaled using standardization.**
* **Target variable (booking\_status) distribution:**
  + **Class 0 (Not Canceled): 24,390 bookings (67%)**
  + **Class 1 (Canceled): 11,885 bookings (33%)**

**2.2 Key Insights from Data**

* **Lead Time: Longer lead times tend to increase the likelihood of cancellations.**
* **Guest Composition (Adults/Children): Bookings with more adults and fewer children were less likely to be canceled.**
* **Booking Timing: Weekend bookings were less likely to be canceled than weekday bookings.**
* **Market Segment and Repeated Guests: The "Online" segment had a higher cancellation rate, while repeated guests were less likely to cancel.**

**2.3 Data Visualizations**

* **Lead Time vs. Cancellation Rate: A positive correlation was observed between longer lead times and higher cancellation rates.**
* **Guest Composition: Bookings with more adults and no children had lower cancellation rates.**
* **Booking Timing: Cancellations were more frequent for weekday bookings.**
* **Market Segment: Cancellations were more common in the "Online" segment, while "Corporate" and "Offline" segments had lower cancellation rates.**

**3. Feature Engineering**

**3.1 Encoding and Transformation**

* **Categorical features such as market\_segment, room\_type\_reserved, and type\_of\_meal\_plan were transformed using One-Hot Encoding.**
* **Numerical features were scaled using StandardScaler to normalize their values.**

**3.2 Correlation Analysis**

* **A heatmap showed that lead\_time, market\_segment, and no\_of\_special\_requests were key features highly correlated with the cancellation status.**

**4. Model Selection and Evaluation**

**4.1 Models Considered**

* **Random Forest Classifier**
* **Logistic Regression**

**4.2 Model Performance**

**Random Forest Classifier**

* **Best Parameters (After Hyperparameter Tuning using GridSearchCV):**
  + **bootstrap: True**
  + **max\_depth: 15**
  + **max\_features: sqrt**
  + **min\_samples\_leaf: 2**
  + **min\_samples\_split: 5**
  + **n\_estimators: 300**
* **Final Training Accuracy: 89.38%**
* **Final Test Accuracy: 87.71%**

**Logistic Regression**

* **Training Accuracy: 75.28%**
* **Test Accuracy: 75.34%**

**4.3 Model Evaluation Metrics**

**Random Forest - Test Set Results:**

* **Precision (Class 1 - Canceled): 0.89**
* **Recall (Class 1 - Canceled): 0.94**
* **F1 Score (Class 1 - Canceled): 0.91**
* **Confusion Matrix:**
  + **True Negative: 2,751**
  + **False Positive: 856**
  + **True Positive: 6,795**
  + **False Negative: 481**
* **AUC-ROC: 0.93 (Training), 0.88 (Test)**

**Logistic Regression - Test Set Results:**

* **Precision (Class 1 - Canceled): 0.77**
* **Recall (Class 1 - Canceled): 0.77**
* **F1 Score (Class 1 - Canceled): 0.77**
* **AUC-ROC: 0.75**

**4.4 Model Comparison**

* **Random Forest performed better across all metrics compared to Logistic Regression, particularly in terms of recall and F1-score for predicting cancellations.**
* **Random Forest is the chosen model due to its superior performance, especially in handling imbalanced data.**

**5. Hyperparameter Tuning**

* **GridSearchCV was used to optimize the Random Forest hyperparameters. The final model configuration was selected based on cross-validation results, ensuring a good balance between training and test accuracy to reduce overfitting.**

**6. Conclusion**

* **The Random Forest model with tuned hyperparameters achieved an accuracy of 87.71% on the test set, making it the best choice for predicting hotel reservation cancellations.**
* **This model provides actionable insights for hotel management, helping them predict cancellations based on booking characteristics like lead time, guest composition, and market segment.**

**Top of Form**

**Bottom of Form**