Predicting Holiday Bookings Using Random Forest

Data Analysis, Model Building, and Key Insights

Problem Statement

- Customers are more empowered due to easy access to information.

- Airlines need to be **proactive** in targeting customers before they travel.

- Task: Use **customer booking data** to build a predictive model for holiday purchases.

Objective

- **Goal:** Predict whether a customer will book a holiday.
- **Approach:**
- Prepare & clean data
- Train predictive models
- Evaluate performance
- Identify most influential variables

Data Understanding

- Dataset: customer_booking.csv
- Contains customer details, booking history, and target variable.
- **Target:** booking_status (Booked / Not Booked)
- Steps performed:
- Missing value check
- Feature selection
- Categorical encoding
- Scaling (for some models)

Model Used

- **Random Forest Classifier**
- Handles mixed data types well
- Resistant to overfitting
- Provides feature importance
- **Hyperparameter tuning:** GridSearchCV
- Parameters tuned: $n\_estimators$, max_depth, $min\_samples\_split$

Model Evaluation

- **Confusion Matrix:**
- True Negatives (TN): **6027** – Correctly predicted Not Booked
- False Positives (FP): **2493** – Predicted Booked but actually Not Booked
- False Negatives (FN): **420** – Predicted Not Booked but actually Booked
- True Positives (TP): **1060** – Correctly predicted Booked

Metrics

- **Accuracy:** (TP + TN) / Total Predictions
- **Precision (Booked):** TP / (TP + FP)
- **Recall (Booked):** TP / (TP + FN)
- **Observation:**
  - High TN count → model good at predicting Not Booked
  - FP is high → need better control to avoid over-predicting Booked

Interpretation of Results

- Many customers predicted as "Booked" did not actually book → Possible overfitting to booking signals.

- False negatives are relatively low, meaning the model is decent at catching actual bookings.

- Possible reasons for FP:
  - Overlapping features between booked & non-booked customers
  - Imbalance in target variable

Improvements Applied

- Hyperparameter tuning with GridSearchCV

- Tried different decision thresholds to reduce FP

- Balanced data using class weights

- Feature engineering to improve predictive power

Conclusion

- Predictive modelling can help airlines target likely bookers in advance.

- Random Forest is effective but needs tuning to control FP.

- Data quality and feature engineering are key for improved results.