Task 1: Web scraping to gain company insights

By: Srinidhi Manikantan

ANALYSIS OF BRITISH AIRWAYS ONLINE REVIEWS

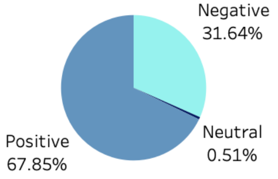
BRITISH AIRWAYS

Percentage of Reviews from different travellers

Economy Class	54.51%
Business Class	30.69%
Premium Economy	9.42%
First Class	5.38%

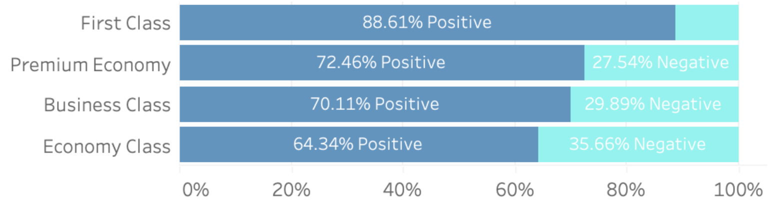
Travellers from Economy Class are more likely to leave a review; More than half the reviews received came from Economy Class.

Sentiment Analysis of Online Reviews



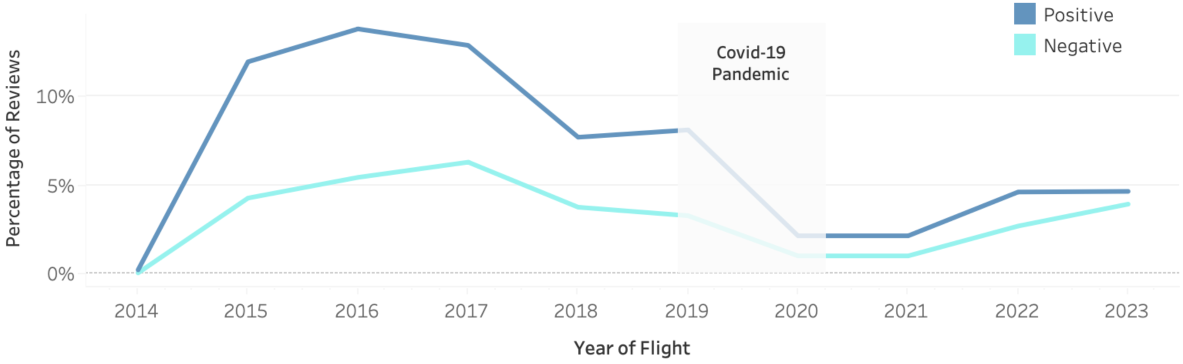
More than two-thirds of the online reviews posted by travellers were positive.

Polarity of Reviews based on Traveller Type



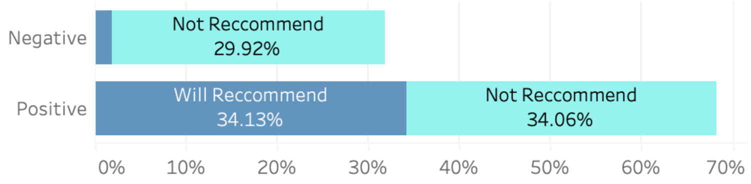
Travellers from First Class are more likey to leave a positive review, while travellers from Economy Class are more likely to leave a negative review

Time-Series Analysis of the Reviews Received



The highest percentage of positive reviews was reported in 2016, whereas the highest percentage of negative reviews was reported in 2017. Due to the Covid-19 pandemic, there is a significant drop in both positive and negative reviews between the period of 2019 to 2020. However, the percentage of reviews have been slowly increasing post-pandemic. It is also important to note that for all the years, the percentage of positive reviews have always been higher than the negative reviews.

Recommendation based on review type



Overall, only 35% of the reviewers voted that they would recommend British Airways to others. Doing a breakdown by review type further revealed that half of the travellers that left a more positive review still chose to not recommend Britisth Airways.

Topic Modeling to find common topics among Reviews Received

Topics	Common Words Identified	Possible Topics Discussed
Topic 1	Seat, Cabin, Food, Crew, Good, Service, Fly, Economy, Time, Drink	In-flight Service and Comfort
Topic 2	Book, Try, Call, Travel, Could, Wait, Tell, Ticket, Customer, Pay	Flight Booking and Payment
Topic 3	Get, Check, Would, Delay, Go, Staff, Time, Service, Hours, Bag	Flight Delays or Baggage Issues
Topic 4	Class, Business, Club, Lounge, First, Excellent, Europe, World, Product, Room	Airport Services (for First/Business Class)

Word Cloud of Most Frequently Used Words

Task 2: Predicting customer buying behaviour

By: Srinidhi Manikantan

UTILIZING MACHINE LEARNING TO PREDICT CUSTOMER BEHAVIOUR

BRITISH AIRWAYS

Exploratory Data Analysis

Original Dataset: *9* numerical + *5* categorical variables, *5000* entries

Checked for missing values, none found

Checked for duplicate values, 719 entries removed

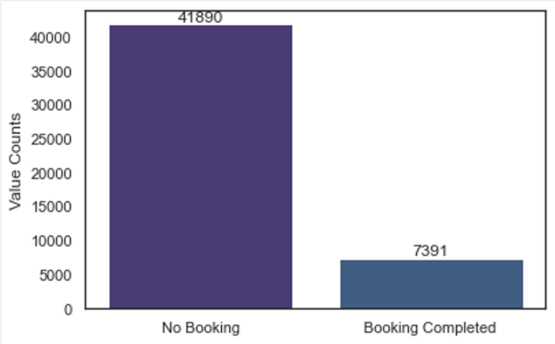
Checked for unique/low variance/correlated variables, none found

Hot-encoded 5 categorical variables to numerical variables

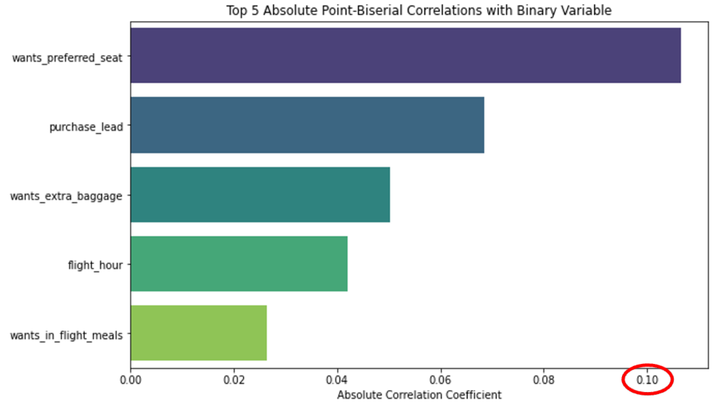
Cleaned Dataset: *14* numerical variables, *49,281* entries

Distribution of Response Variable

In the cleaned dataset, 85% of the majority class represents "No Booking", while the remaining 15% belongs to "Booking Completed". This represents an imbalanced dataset, where one class significantly outnumbers the other.



Correlation of Predictor Variables with Response Variable

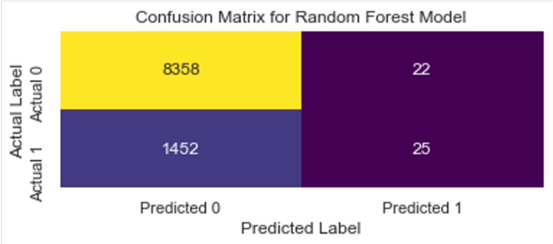


Since the target variable is binary, **Point-Biserial Correlation** was used as a correlation metric.

There was **no significant correlation** identified, and the highest correlation was only **0.10**, proving that there is no to very weak associations between the predictor and response variables.

Assessment of ML Models

Models Used	Accuracy with Cross-Validation	Accuracy with Hyperparameter Tuning	F1 Score with Hyperparameter Tuning
Logistic (all penalties)	0.843	0.845	0.0142
Random Forest	0.688	0.849	0.0512
Extra-Trees	0.806	0.850	0.0173
Grad Boost	0.682	0.851	0.0443



Cross-validation was conducted among 8 different base Machine Learning models (Logistic, RandomForest, ExtraTrees, XgBoost, Decision Tree, GradBoost, CatBoost and LightBGM).

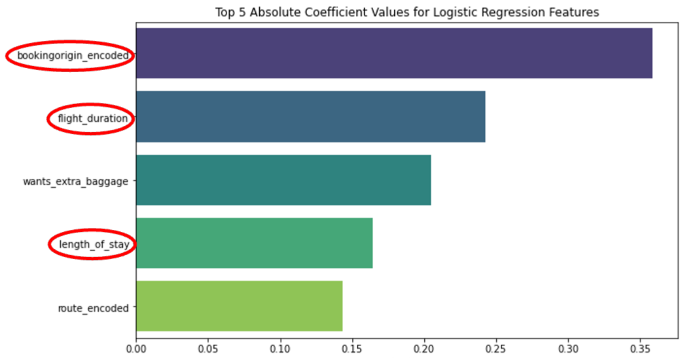
Top 4 models with highest scores were chosen and **hyperparameter tuning** was performed to further improve the model.

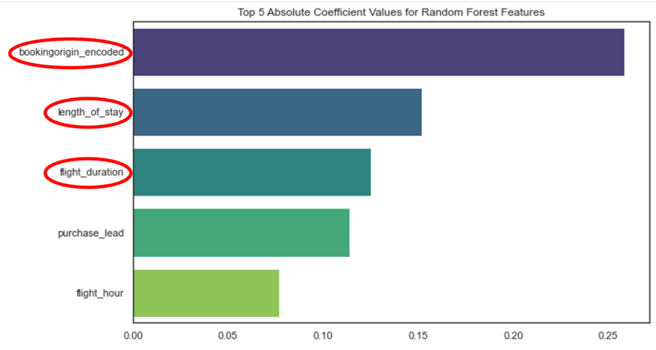
Since there is a significant class imbalance, the model can achieve high accuracy by simply predicting the majority class most of the time, but this is not useful in practice. Hence, **F1 score would be more informative**, and meaningful since it balances both precision and recall in a single metric.

Random Forest and Grad Boost were the highest performing models. However, the F1 score is significantly low for both models (~0.05) and may not predict the customer buying behaviour accurately in real life.

Finding Important Predictor Variables

Using **Feature Importance** from Logistic Regression and Random Forest, **top 3 common features** identified are *Booking Origin, Length of Stay and Flight Duration*.





Limitations and Future Directions

- **Insufficient Data:** The dataset consists of only 5000 entries, which might be insufficient to capture the complexity and diversity of customer behaviours, particularly for a large-scale operation like British Airways with over 40 million customers annually. This may lead to suboptimal model generalization
  - ➢ Consider additional data collection (entries) and more consumer-centric feature variables (e.g. demographic characteristics) to improve model predictability
- **Imbalanced Data**: Imbalanced datasets can lead to biased model training, where the model favours the majority class and may not generalize well to the minority class. This may lead to lower precision, recall, and F1 score for that class.
  - ➢ Consider collecting more samples from the underrepresented minority class.