Model Performance Report: Customer Attrition Prediction

Team 13

Purpose of the Models

Predict customer attrition using:

- Linear Regression
  - Used to predict a continuous outcome variable based on one or more predictor variables by fitting a linear relationship (line) to the data, which minimizes the difference between the observed and predicted values.
- Logistic Regression
  - Used to predict a binary outcome (0 or 1) based on one or more predictor variables by modeling the probability of the outcome using a logistic function, which outputs values between 0 and 1.

Data Cleaning

- Handle missing data/value
- Remove string from column Cost, Car Rental Addon, and Hotel Addon
- Add respective column in USD

```
# Function to extract numeric value and convert to USD
def convert_to_usd(value, exchange_rates):
    currency_code = value[-3:]  # Assuming the last 3 characters are the currency code
    amount = float(value[:-4])  # Convert everything except the last 3 characters and a space to a float
    # Convert to USD using the provided exchange rate, default to 1 if currency not found
    amount_in_usd = amount * exchange_rates.get(currency_code, 1)
    return currency_code, amount_in_usd
```

- Removing outliers
- Convert to Binary

Feature Engineering

- Tenure of the Frequent Flyer

```
# Calculate tenure as the difference in days, months, or years
today = datetime.today()
df2['Tenure (Years)'] = (today - df2['Join Date']).dt.days / 365
```

- Average spend yearly per Frequent Flyer for the last 2 years

```
# Group by Frequent Flier Number to calculate total spend over the two years (2024-2025)
total_spend = df2_filtered.groupby('Frequent Flier Number')['Total Spend in USD'].sum().reset_index()
total_spend.columns = ['Frequent Flier Number', 'Total Spend in USD (2024-2025)']
```

- Merge two dataframes into one

```
# Select necessary columns from df1
df1_relevant = df1[['Frequent Flier Number', 'Inquiry Type', 'Lounge Used?', 'Planned Snack?'

# Merge df1 and df2 on 'Frequent Flier Number'
merged_df = pd.merge(df1_relevant, df2, on='Frequent Flier Number', how='inner')
```

- One-Hot Encoding for Categorical data (Inquiry Type)

Linear Regression: Price Prediction

- Linear regression are computationally efficient, making them quick to train, even on larger datasets.
- Key Metrics:
  - Mean Squared Error (MSE)
  - Mean Absolute Error (MAE)
  - R^2 Score
- Features used:

'Lounge Used?', 'Planned Snack?', 'Additional Snack?', 'Flight Delayed?', 'Tenure (Years)', 'Average Yearly Spend in USD', 'Inquiry_Cancel Flight', 'Inquiry_Flight Deal', 'Inquiry_Flight Status', 'Inquiry_New Flight'

Train the Model

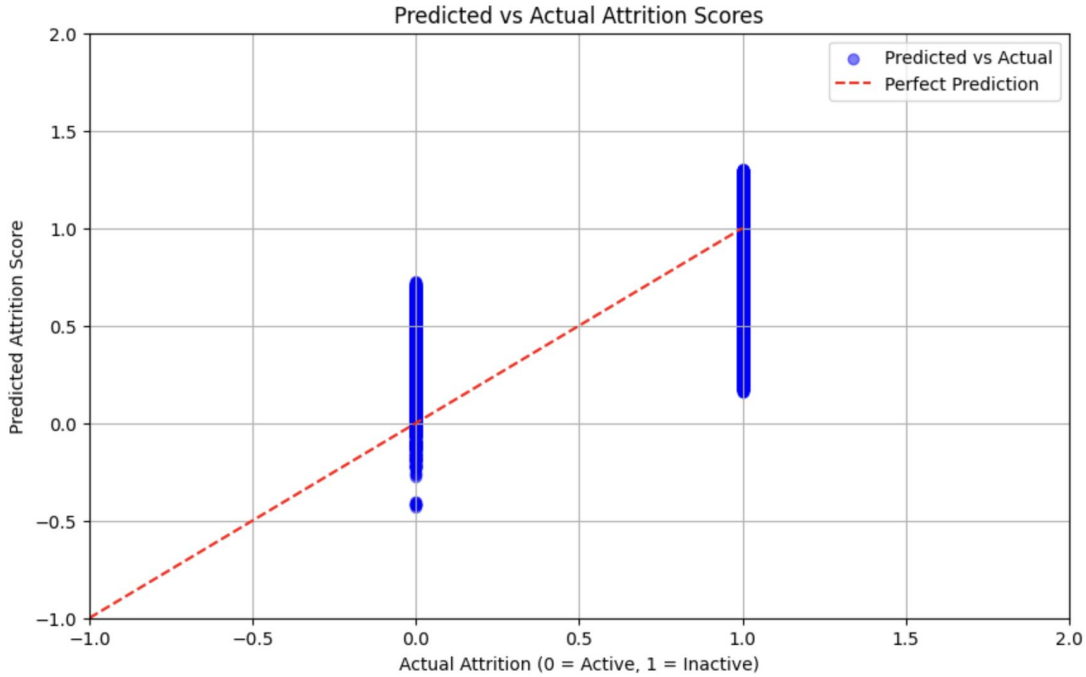
With test data 20%

Key metrics:

MSE	0.0823
MAE	0.4867
R^2	0.48

Higher R^2 implies better prediction of the model.

```
Mean Squared Error: 0.11931311530484429
Mean Absolute Error: 0.2843934263283781
R-squared: 0.4867306037644137
Predicted Attrition Scores:
 [0.4285706  0.59128907 1.04334533 ... 0.94843814 0.76782185 0.37621122]
```



Logistic Regression: Delay Prediction

- Logistic regression provides probabilities for the predicted class and tends to perform well even when the assumptions are not fully met.
- Key Metrics:
  - Accuracy
  - Confusion Matrix
  - Classification Report
- Features used:

'Lounge Used?', 'Planned Snack?', 'Additional Snack?', 'Flight Delayed?', 'Tenure (Years)', 'Average Yearly Spend in USD', 'Inquiry_Cancel Flight', 'Inquiry_Flight Deal', 'Inquiry_Flight Status', 'Inquiry_New Flight'

Train the Model

- With test data 20%, the model demonstrate good performance with accuracy approximately is 84.8%
- True: Customer **will not** remain a frequent flyer
- False: Customer **will** remain a frequent flyer

```
Attrition
True     43563
False    25648
Name: count, dtype: int64
Accuracy: 0.8480820631366034
Confusion Matrix:
 [[4461  627]
 [1476 7279]]
Classification Report:
              precision    recall  f1-score   support

       False       0.75      0.88      0.81      5088
        True       0.92      0.83      0.87      8755

    accuracy                           0.85     13843
   macro avg       0.84      0.85      0.84     13843
weighted avg       0.86      0.85      0.85     13843
```

Test Case

Based on a single frequent flyer number: 7234617746

```
Thresholds Used - Tenure: 18.660344836984667  | Yearly Spend: 12781.71105

Linear Regression Prediction for Debbie Spears (Attrition Score): 0.10517235352363663
R-squared: 0.4867306037644137
Based on the linear prediction, Debbie Spears is likely to remain a frequent flyer.

Logistic Regression Class Prediction for Debbie Spears (Attrition): False
Logistic Regression Probability Prediction for Debbie Spears (Inactive): 0.020566894506464917
Accuracy: 0.8480820631366034
Based on the predictions, Debbie Spears is likely to remain a frequent flyer.
```

Based on all data:

```
Count of Frequent Flyers by Status (Linear Regression):
                                        Status  Count
0  not likely to remain a frequent flyer   1230
1      likely to remain a frequent flyer    770
```

Recommendations

- Rank customer attrition based on key features, such as flight counts, spending habits, and inquiry types.
- Feature Engineering
  - Customer engagement metrics
    - Loyalty program
    - Miles earned or redeemed
  - Customer feedback
- Model Improvement suggestions
  - Balance the dataset using resampling techniques
  - Use regularization
  - Use other models such as Random Forest, Neural Network