Flight Prices Prediction

Anh Nguyen, Fiona Xu, Sadichchha Maharjan

Topic Introduction & Previous Works

- Topic: predicting flight prices using features such as departure time, distance travelled, ticket class, etc.
- Many solutions with traditional machine learning algorithms (KNN, Random Forest, Decision Tree)
- Papers utilizing deep learning have shown better performance because of the ability to better capture complex patterns and temporal dependencies
- Our project: Comparing the performance of multiple models for this task

Dataset

- https://www.kaggle.com/datasets/dilwong/flightprices
- Almost 6 million US flights between 2022-04-17 and 2022-11-11
- Includes one-way flights from major U.S. airports like ATL, JFK, LAX, SFO, and more
- 27 data attributes (destination, airline, flight duration, ticket price, etc.)

Data Preprocessing

- Dropping unnecessary columns (flight id)
- Creating new columns (isCoach, days_until_flight, isHoliday, isWeekend, etc.)
- Removing null values
- Scaling for numerical features, one-hot encoding for categorical features
- Train-validation-test split

Baseline model: Linear Regression

- Fit the model using Ordinary Least Squares regression (OLS)
- Can see model summary and significant features
- Results:

```
MSE: 11037.3771
MAE: 76.6251
R²:  0.2728
```

Baseline model: XGBoost

- N-estimators = 1000
- Early stopping 50
- Trained for 999 boosting rounds
- Results:
- MSE: 5304.88
- MAE: 43.87
- R^2:0.6505

Deep Learning model: LSTM

- two LSTM layers with 50 units each
- applies dropout to prevent overfitting
- Early stopping with patience at 3
- Total epochs set 50, actually trained 5: model is getting worse as it trains

MSE: 19966.8
MAE: 109.25
R^2:-0.3155

Modified and tested out lstm layers with scaled target values, results was worse: might be due to problems with input values

Deep Learning model: GRU

```
model = Sequential([
    GRU(32, return_sequences=True, input_shape=(1, train_gen.X.shape[1])),
    Dropout(0.3),
    GRU(16),
    BatchNormalization(),
    Dropout(0.3),
    Dense(1)
])
```

- **GRU model:** Captures long-term dependencies efficiently
- Target variables were scaled using (0 to 1) to improve learning stability
- **Results:**

    MSE: 8900.3379, MAE: 73.4945, R^2: 0.1073

    MSE: 7108.9492, MAE: 63.8582, R^2: 0.2870

Discussion

- Challenges: Complex, dynamic pricing patterns not fully captured
- Large, complex dataset:
  - High dimensionality with mixed categorical & numerical features
- Dynamic pricing:
  - Flight prices change frequently based on many external factors
- **Next Steps:**
  - Tune hyperparameters and run more epochs to improve model performance

Thank you for listening!