Flight Prices Prediction Milestone 2

Anh Nguyen, Fiona Xu, Sadichchha Maharjan

Topic and Milestone 1 recap

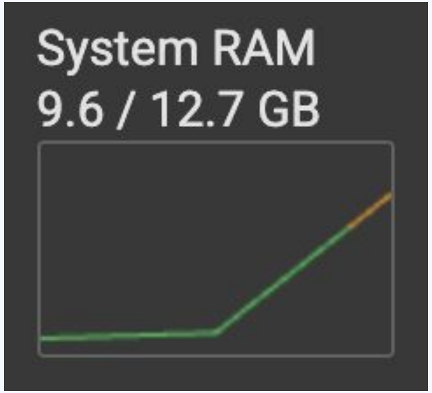
- Topic: predicting flight prices using features such as departure time, distance travelled, ticket class, etc.

- Milestone 1 : Data exploration, cleaning, preprocessing completed, we now have train-test-validation sets and X, Y ready for modeling

- Milestone 1 challenge: very big and complicated dataset (a lot of cleaning, encoding, time needed, significantly slows down progress and will need good feature selection)

Baseline model: ARIMA

- **Autoregressive Integrated Moving Average:** well-suited for time series data, capturing trends, seasonality, and autocorrelations
- **Approach:**
  - Use daily flight prices and applied ARIMA to model temporal trends in flight fares
- **Roadblocks:**
  - Requires a Pandas DataFrame — conversion from PySpark was memory-intensive



System RAM
9.6 / 12.7 GB

Baseline model: XGBoost

- **eXtreme Gradient Boosting**
- Supervised learning
- Gradient boosting: builds a series of decision trees where each tree tries to correct the errors made by the previous ones (an ensemble learning algorithm)
- Compared to Random Forest: handles trees sequentially, minimize bias and underfitting-result sums all the trees
- No time assumptions good for our data: multivariate, nonlinear patterns
- **Cons-overfitting**
- Includes regularization to prevent overfitting

```
from xgboost.spark import SparkXGBRegressor
```

https://xgboost.readthedocs.io/en/stable/tutorials/spark_estimator.html

https://www.kaggle.com/code/robikscube/tutorial-time-series-forecasting-with-xgboost
https://www.kaggle.com/code/robikscube/pt2-time-series-forecasting-with-xgboost
https://ieeexplore.ieee.org/document/9793411

Why 2 models

Compare time series model with more complex ML model

Understand features better: time features vs. other features

Deep Learning model: RNN & Evaluation

- Start with simple RNN, then try LSTM, GRU

https://ieeexplore.ieee.org/document/10900524 (article implemented RNN, GRU, LSTM for flight price prediction)

- Evaluation Metrics: R squared, RMSE
- Test set, cross validation, external datasets

Thank you for listening!