Unlocking Insights: Leveraging Predictive Modeling to Understand the Impact of Weather Conditions on Delivery Times

Arun Bhatia
arun.bhatia@aalto.fi

Introduction: Overview of the Task and Objectives

Problem statement

- The logistics industry faces delays and disruptions due to weather, impacting service quality
- The challenge is figuring out how different weather conditions affect delivery times
- We need to optimize delivery services, ensuring reliability and keeping customers happy

Objectives

1. Understanding the impact of weather conditions
2. The primary goal is to develop a predictive model that unveils the relationship between weather conditions and delivery times
3. Address the challenges posed by varying weather conditions on delivery times

Data Preprocessing and Feature Engineering

- We observed 277 missing values in each feature (CLOUD_COVERAGE, TEMPERATURE, WIND_SPEED)
  - Missing values were in the same dataset entries /on the same line -> Remove these lines
- Creating interaction features helps capture potential non-linear relationships among weather features, enhancing our model's ability to recognize complex patterns
- At the end: Total 13 features with 3 interaction features and 18429 data points

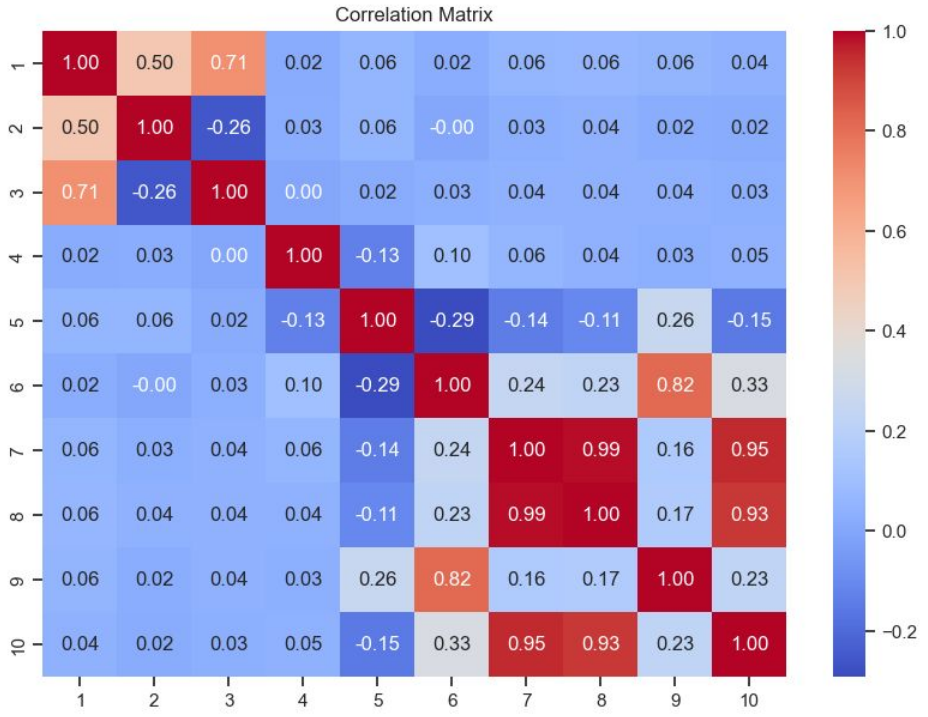
```
Data columns (total 13 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   TIMESTAMP                                       18429 non-null  datetime64[ns]
 1   ACTUAL_DELIVERY_MINUTES                         18429 non-null  int64
 2   ESTIMATED_DELIVERY_MINUTES                      18429 non-null  int64
 3   ACTUAL_DELIVERY_MINUTES - ESTIMATED_DELIVERY_MINUTES  18429 non-null  int64
 4   CLOUD_COVERAGE                                  18429 non-null  float64
 5   TEMPERATURE                                     18429 non-null  float64
 6   WIND_SPEED                                      18429 non-null  float64
 7   PRECIPITATION                                   18429 non-null  float64
 8   hour_of_day                                     18429 non-null  int32
 9   day_of_week                                     18429 non-null  int32
 10  temperature_precipitation_interaction           18429 non-null  float64
 11  temperature_wind_speed_interaction              18429 non-null  float64
 12  precipitation_wind_speed_interaction            18429 non-null  float64
```

Modeling Approach

The correlation matrix doesn't show any significant correlation between delivery time features and weather condition features

This is likely because the relationship between delivery times and weather conditions may involve more complexity, possibly non-linear patterns

Based on the correlation matrix, we'll use the following weather-related features to build our model: 'CLOUD_COVERAGE', 'TEMPERATURE', 'WIND_SPEED', 'PRECIPITATION', 'temperature_precipitation_interaction', 'temperature_wind_speed_interaction', 'precipitation_wind_speed_interaction'.




```
Feature Numbers:
1: ACTUAL_DELIVERY_MINUTES
2: ESTIMATED_DELIVERY_MINUTES
3: ACTUAL_DELIVERY_MINUTES - ESTIMATED_DELIVERY_MINUTES
4: CLOUD_COVERAGE
5: TEMPERATURE
6: WIND_SPEED
7: PRECIPITATION
8: temperature_precipitation_interaction
9: temperature_wind_speed_interaction
10: precipitation_wind_speed_interaction
```

Predictive Machine Learning Model & Validation

- We know that the relationship between delivery times and weather conditions isn't straightforward. There seem to be non-linear patterns or other factors
- We choose to use Random Forest, an ensemble learning method
  - Random Forest is known for capturing complex, non-linear relationships within data
  - Random Forest is robust to noisy data and outliers. In real-world scenarios, datasets may contain irregularities
  - By combining multiple decision trees, Random Forest reduces the risk of overfitting and enhances generalization
  - Random Forest often performs well in predictive tasks, making it a reliable choice here
  - Random Forest model provides insights into feature importance, allowing us to understand the impact of different features on delivery times

- We will also use k-fold cross-validation which allows us to train and test our model on different subsets of the data. This ensures a more robust estimate of the model's performance. In our case, we use k=5, meaning the dataset is divided into 5 folds. The model is trained and evaluated 5 times, each time using a different fold as the test set and the training set

Model Evaluation & Results (1/2)

```
Cross-Validation Scores:
MAE: [8.83908765 8.84520416 9.20889121 8.52125126 8.17946002]
MSE: [116.51027265 116.79929489 128.8684905  108.89746616  99.74637727]
R2 Score: [-0.21083235 -0.18000506 -0.27363511 -0.08088517 -0.12982585]
MAPE: 25.91825106392379%

Mean Absolute Error (MAE) across folds: 8.718778861120375
Mean Squared Error (MSE) across folds: 114.16438029472593
R2 Score across folds: -0.17503670750365125
```

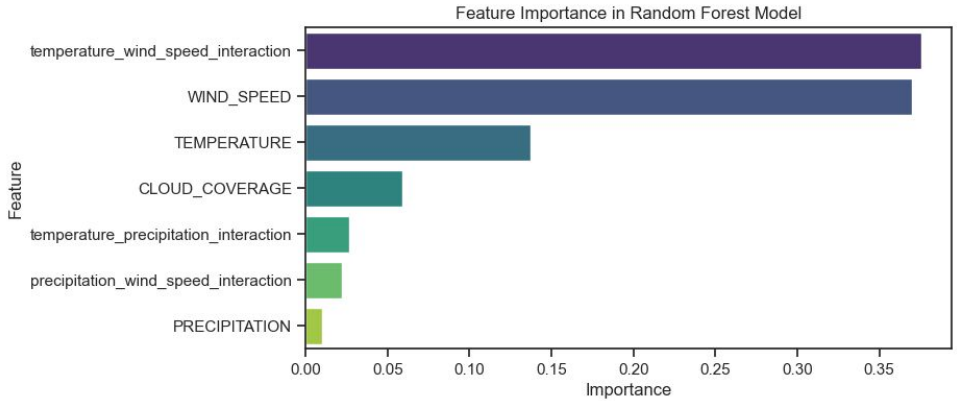
Key insights for the cross-validation scores:

- The model demonstrates reasonable predictive ability, with an average error of 8.72 minutes = 8 m 43 s
- R2 scores are negative across all folds, indicating suboptimal model performance. A mean R2 score of approximately -0.17 suggests that the model falls short of capturing the variance in the data
- MSE provides a more sensitive measure to outliers, MSE values across the folds show some variability, indicating that the our model's performance is not consistent across different subsets of the data
- MAPE tell us that on average, predictions differ by around 25.92% from the actual delivery times

Model Evaluation & Results (2/2)

Visualizing feature importance scores can helps us understand which features are contributing the most to the model's prediction

- Temperature and wind speed interaction stands out as the most crucial factor, suggesting a nuanced relationship. Wind speed alone significantly affects delivery times.
- Precipitation alone has the least influence on delivery times.



The evaluation of the feature 'ESTIMATED_DELIVERY_MINUTES' in the dataset reveals better performance compared to our predictive model. It demonstrates lower Mean Absolute Error (MAE), Mean Squared Error (MSE), and a higher R2 Score, indicating more accurate estimates

```
ESTIMATED_DELIVERY_MINUTES evalution for comparison:

MAE: 7.253730533398448
MSE: 82.10033100005427
R2 Score: 0.18240710822980433
```

Limitations & Further Development

Limitations

- Data Limitation: The model's predictions heavily depend on the quality and completeness of the dataset
- Simplification: The model assumes a linear relationship between some weather conditions and delivery times. There is likely more complex, non-linear patterns between other interaction features than the 3 created

Further Development

- External Factors: The model focuses on weather conditions, but external factors like traffic, unexpected events, or temporal elements should also influence delivery times and these need to be added to future model
- Use more advanced modelling techniques
  - Fine-tuning the hyperparameters
  - Having more weather-related features
  - Experimenting with different machine-learning algorithms