Understanding Weather Impact on Courier Availability

Muhammad Hassan Shafiq

January 30, 2024

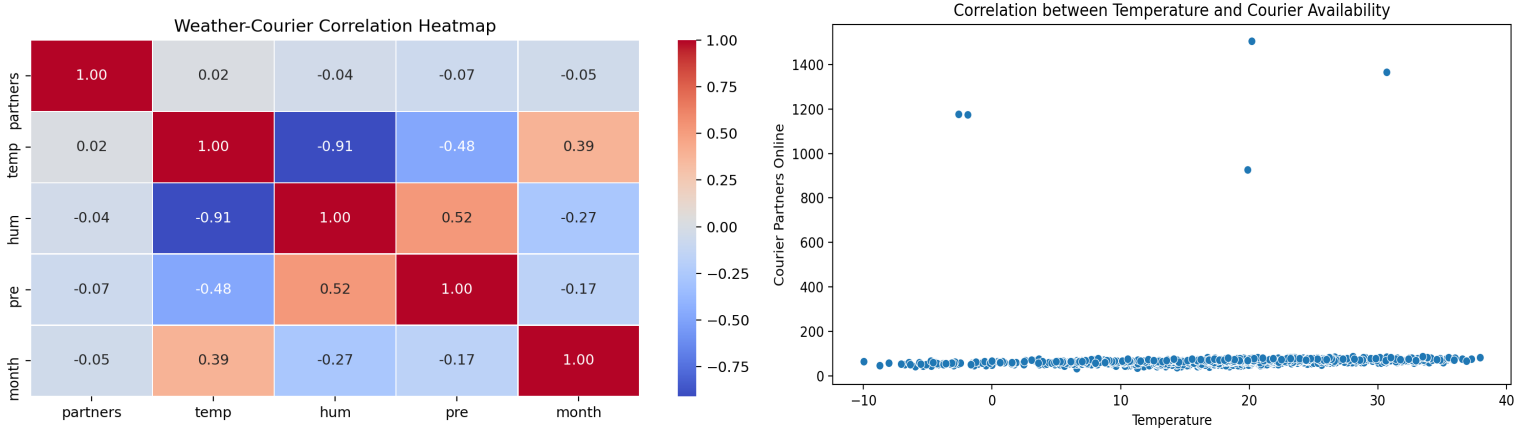
Introduction

- The use case is the analysis for understanding the impact of weather conditions on courier availability.
- The analysis explores the relationship between weather factors such as temperature, relative humidity, and precipitation, and the number of courier partners available on a given day.
- The target audience for the presentation is potential future peer Data Scientists with relevant business knowledge.

Data Preprocessing

- Imputed missing values in **temperature and precipitation columns**.
- Instead of finding the mean value of the entire temperature column and imputing it, mean values were computed for each individual month.
- This approach was chosen due to the significant variation in temperature experienced in Finland, particularly during winter months.
- By using the mean value of a particular month, the impact of lower or higher temperature values specific to each month could be better accounted for.
- The decision was based on the assumption that temperature variations follow a noticeable trend over the course of a month, making month-specific means a more accurate representation.

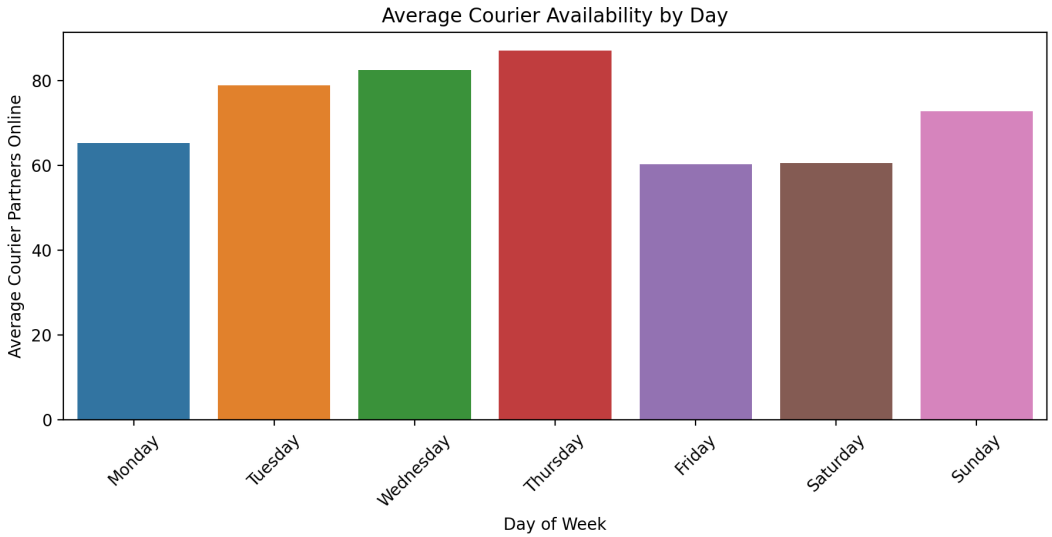
Correlation Analysis



- **Developed a heat map** to analyze the relationship between various factors and courier availability.
  - Found a positive correlation between temperature and the number of courier partners, suggesting that **warmer weather may lead to increased courier availability**.
  - Observed a negative correlation between precipitation and courier availability, indicating that **rainy days may impact the availability of courier partners**.
- Plotted the correlation between temperature and courier availability separately, confirming the earlier hypothesis and also found out some outliers in the data.

Data Analysis Insights

- Identified specific dates with potentially unusual courier availability, possibly due to events, holidays, or promotional activities.
  - Omitted these values to improve model accuracy.
- Analyzed the average courier availability by day:



  - Found that partners prefer to work more on Tuesday, Wednesday, Thursday, and Sunday, with Thursday having the highest availability.

Model Development

I divided the data into training and testing sets following the 80 to 20 ratio for better results and used 2 different models.

Linear Regression Model:

- Chosen Linear Regression due to its simplicity and interpretability, making it suitable for understanding the linear relationship between weather conditions and courier availability.
- Benefits:
  - **Interpretability**: Linear Regression provides clear insights into how changes in weather variables impact courier availability, facilitating easy interpretation of results.
  - **Ease of Implementation**: The model is straightforward to implement and understand, making it accessible even to non-experts.
  - **Baseline Performance**: Serves as a baseline model for comparison with more complex algorithms, helping to gauge the incremental improvement achieved by advanced techniques.

Model Development Continued

Random Forest Regression Model:

- Chosen Random Forest Regression for its ability to handle nonlinear relationships and interactions between features, making it robust for complex datasets like weather conditions.
- Benefits:
  - **Nonlinearity Handling**: Random Forest Regression can capture nonlinear relationships between weather variables and courier availability, providing flexibility in modeling.
  - **Feature Importance**: The model offers insights into the importance of different weather features, helping to identify key drivers of courier availability.
  - **Ensemble Learning**: Utilizes ensemble learning techniques to combine multiple decision trees, reducing overfitting and improving generalization performance.
  - **Robustness**: Random Forest is less sensitive to outliers and noise in the data compared to linear models, enhancing its robustness and stability.

Metrics for Evaluation and Future Steps

- **Linear Regression Model**:
  - RMSE: 7.51, MAE: 5.69, R-squared: 0.32
- **Random Forest Regression Model**:
  - RMSE: 7.63, MAE: 5.90, R-squared: 0.30
- I have used these metrics as they help in explaining the accuracy, precision, and goodness of fit of the models.
- Used Google weather data for January 30th to February 4th and predicted courier partners. More complex models can be developed for further enhancing the performance

Date	Temp	RH	PPTN	Prediction(LR)	Prediction(RF)
2024-01-30	0	0.89	0.00	58.960435	63.116199
2024-01-31	1	0.93	0.05	59.530217	62.713709
2024-02-01	1	0.75	0.35	58.287239	61.927979
2024-02-02	-1	0.84	0.00	58.273508	62.602967
2024-02-03	1	0.76	0.50	58.115841	61.407785
2024-02-04	-2	0.79	0.00	57.586581	63.616557