PREDICTIVE MODELING FOR WOLT

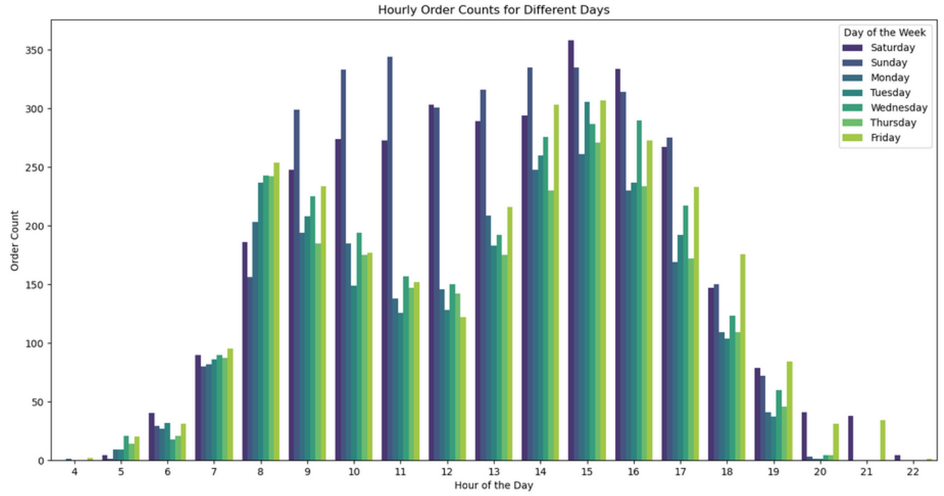
DATA SCIENCE INTERNSHIP ASSIGNMENT 2024

PEPPI KIRKKOMÄKI

DATASET OVERVIEW

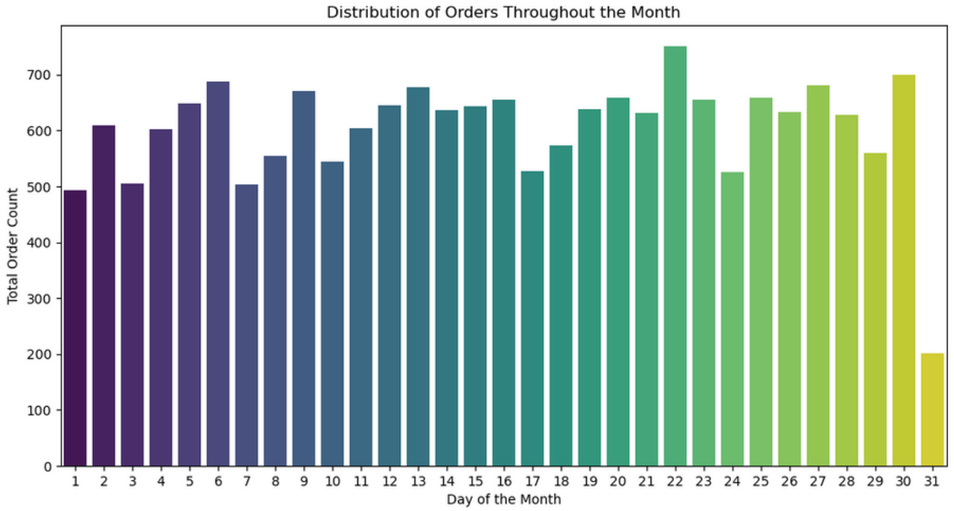
Order Distribution Per Month

There is a clear spike in the days of the week for weekend orders at breakfast and lunch and another spike at dinner compared to weekdays. There is clearly also a high level of ordering during the weekday, but the number of orders increases during the weekend at lunch and dinner times. This is certainly reflected in the fact that people also order more for breakfast and brunch at the weekend than on weekdays.



Order Distribution Per Month

Orders are regularly placed regardless of the date of the month. For example, common pay days such as the first, fifteenth or last day of the month have no obvious change in order numbers.



Order Count and Average Delivery Time

From lunchtime onwards, there is a clear increase in the number of orders. While there is a clear increase in the average delivery time of orders during lunch and dinner. Towards the evening, the average order delivery time decreases significantly.

FEATURE SELECTION AND HANDLING

hour_of_day

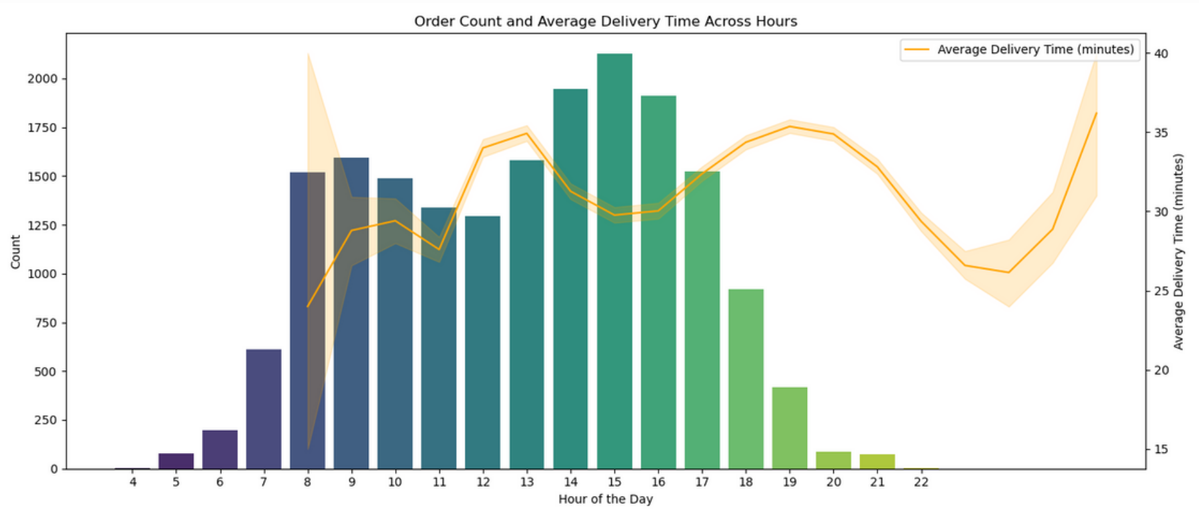
This is an excellent illustration of hourly patterns in the data.

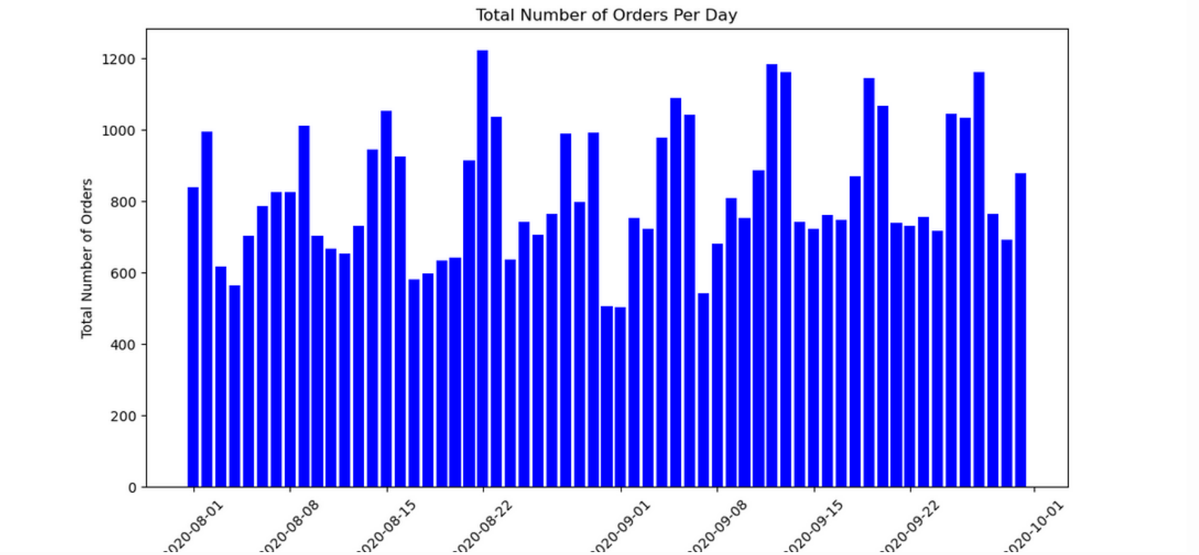
day_of_week

This is a good illustration of how the day of the week plays a role in order numbers.

Handling Missing Values

- Missing values were addressed using mean imputation.
- Imputed missing values in selected features to maintain data completeness.
- Mean imputation involves replacing missing values with the mean of the observed values for that feature.
- This method was chosen for its simplicity and effectiveness in maintaining feature integrity, ensuring a representative dataset for modeling.





MODELING APPROACH AND TRAINING

CHOSEN MODELS

Linear Regression

Opted for Linear Regression due to its simplicity and interpretability. Suitable for capturing linear relationships between features and the target variable.

Random Forest Regression

Selected Random Forest Regression for its ability to handle non-linearities and capture complex relationships within the data.

TRAINING STRATEGY

Utilized TimeSeriesSplit for training and validation.

- Preserves temporal ordering, crucial for modeling time-dependent trends.
- Effectively handles potential variations in ordering patterns over time.

```python
# Train the Linear Regression model
lr_model = LinearRegression()
lr_model.fit(X_train, y_train)

# Make predictions on the test set
lr_predictions = lr_model.predict(X_test)

# Evaluate the model
lr_mse = mean_squared_error(y_test, lr_predictions)
print(f"Linear Regression Mean Squared Error: {lr_mse}")
lr_mae = mean_absolute_error(y_test, lr_predictions)
print(f"Linear Regression Mean Absolute Error: {lr_mae}")
```

```
Linear Regression Mean Squared Error: 79.46777263276682
Linear Regression Mean Absolute Error: 7.090048023290165
```

```python
# Train the Random Forest Regression model
rf_model = RandomForestRegressor()
rf_model.fit(X_train, y_train)

# Make predictions on the test set
rf_predictions = rf_model.predict(X_test)

# Evaluate the model
rf_mse = mean_squared_error(y_test, rf_predictions)
print(f"Random Forest Regression Mean Squared Error: {rf_mse}")
rf_mae = mean_absolute_error(y_test, rf_predictions)
print(f"Random Forest Regression Mean Absolute Error: {rf_mae}")
```

```
Random Forest Regression Mean Squared Error: 83.00976025996201
Random Forest Regression Mean Absolute Error: 7.24204400498842
```

EVALUATION METRICS AND RESULTS

SELECTED EVALUATION METRICS

Utilized Mean Squared Error (MSE) and Mean Absolute Error (MAE) to assess model performance.

MODEL PERFORMANCE RESULTS

Linear Regression:

- Mean Squared Error: 79.47
- Mean Absolute Error: 7.09

Random Forest Regression:

- Mean Squared Error: 83.01
- Mean Absolute Error: 7.24

INTERPRETATION OF RESULTS

Linear Regression:

- The MSE of 79.47 indicates the average squared difference between predicted and actual values, while the MAE of 7.09 represents the average absolute difference.
- Despite a relatively high MSE, the MAE suggests that, on average, predictions are within 7.09 units of the actual values.

Random Forest Regression:

- The higher MSE of 83.01 suggests a slightly larger average squared difference compared to Linear Regression. The MAE of 7.24 represents the average absolute difference.

COMPARATIVE ANALYSIS

- The comparative analysis reveals [discuss any patterns, strengths, or areas for improvement].
- Consideration for further model refinement and tuning based on these results.

MODEL COMPARISON AND DEVELOPMENT

COMPARISON OF MODELS

Linear Regression:

- Strengths:
  - Simplicity and interpretability make it easy to understand and communicate results.
  - Fast training and prediction times, suitable for large datasets.
- Weaknesses:
  - Assumes a linear relationship between features and the target, limiting its ability to capture complex non-linear patterns.

Random Forest Regression:

- Strengths:
  - Capable of capturing complex non-linear relationships in the data.
  - Robust against overfitting, providing reliable predictions on diverse datasets.
- Weaknesses:
  - May be computationally expensive and time-consuming for large datasets.
  - Interpretability can be challenging due to the ensemble nature of the model.

MODEL COMPARISON AND DEVELOPMENT

SUGGESTIONS FOR IMPROVEMENT

Linear Regression:

- **Hyperparameter Tuning:** Fine-tune model parameters, such as the number of trees, depth of trees, or learning rate, to optimize performance.
- **Feature Engineering:** Experiment with additional features or transformations to enhance model understanding and prediction accuracy.
- **Ensemble Models:** Combine the strengths of both models through ensemble techniques for improved overall performance.

Considerations for Deployment

- **Scalability:** Assess the scalability of the models for real-world deployment, considering potential increases in data volume.
- **Interpretability:** Balance model complexity with interpretability, ensuring that results can be easily understood by stakeholders.
- **Monitoring and Maintenance:** Establish a plan for continuous monitoring and model maintenance to adapt to evolving data patterns.

RELATING FINDINGS TO WOLT

PREDICTIVE INSIGHTS FOR WOLT

The predictive models unveil actionable insights for Wolt's operations, such as:

- Forecasting daily order volumes to optimize resource allocation.
- Identifying peak ordering times and locations for efficient courier management.
- Anticipating variations in courier availability based on historical data.

OPERATIONAL EFFICIENCY

- These insights contribute to enhancing operational efficiency, ensuring timely deliveries, and optimizing resource utilization.

INSIGHTS GAINED

Throughout this assignment, I've gained valuable experience in:

- Exploratory Data Analysis (EDA) for understanding temporal patterns in order data.
- Feature engineering to extract meaningful information from diverse datasets.
- Implementing and evaluating predictive models to forecast delivery-related metrics.

AND WHO AM I?

MY BACKGROUND

Studies

I am a second year student of IT-Bussines at Turku University of Applied Sciences. I specialise in Artificial Intelligence and Data Engineering.

Work background

I currently work in customer support at Wolt while attending school. Before that, I have a long experience in the retail industry, both in customer service and sales.

Why data at Wolt?

My interest in data grew when I worked in the sales team at Kesko. And as I've always been good at maths I thought working with data would be a good career for me.

I'd like to get ahead in my Wolt career and see new sides of the company I already work for. I've enjoyed my time at Wolt and it would be great to progress into a job in my own field.

THANK YOU FOR READING!