Wolt Data Science Internship 2024

Title: Predicting Venue Popularity for Wolt

Name: Kyriakos Kyriakou

Date: 30.01.2024

Introduction to the Dataset and Task

Dataset Overview:

- **Source:** Wolt's 2020 operational data.
- **Content:**
  - Temporal data – when orders were placed.
  - Order details – items count in the order, actual / estimated delivery times.
  - **Geospatial data** – users' and venues' latitude and longitude.
  - Environmental conditions – cloud coverage, temperature, wind speed, and precipitation when orders were placed.
- **Scope:** Captures the fluctuating dynamics of customer orders in Helsinki.

Objective:

- **Task:** Predict venue popularity tiers (High, Medium, Low) based on geospatial coordinates.

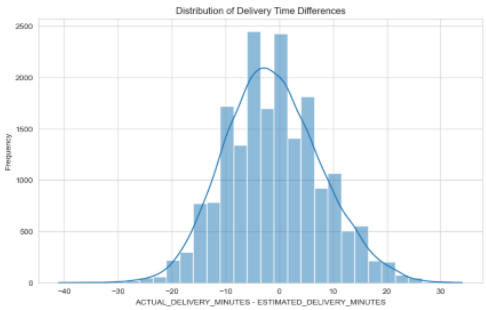
Significance to Wolt:

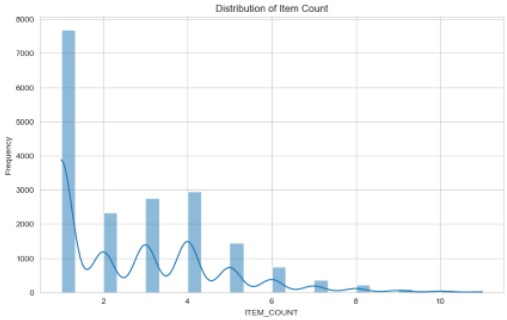
- **Resource Allocation:** Direct resources to high-demand venues to minimize waiting times and improve delivery efficiency.
- **Strategic Insights:** Understand patterns of demand across different areas to guide marketing and partnership strategies.
- **Market Intelligence:** Gauge potential venue success to support decisions for future collaborations.

Data Exploration

Key Statistics:

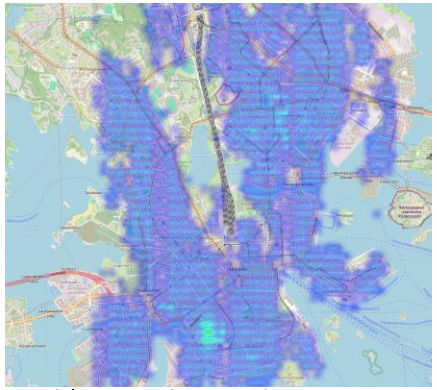
- **Order Volume:** Total number of orders in dataset: 18706.
- **Delivery Times:** Average actual vs estimated delivery times: 1.201 minutes faster.
- **Item Count** average per order: 2.7 items.

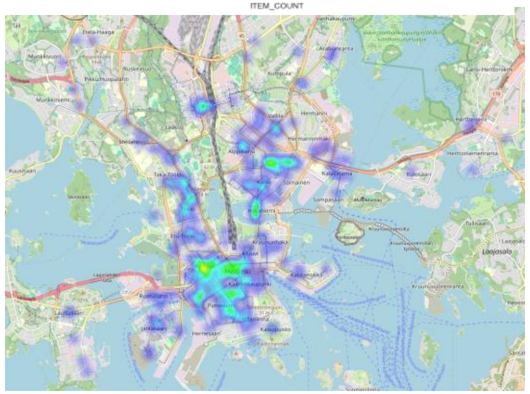




Geospatial Insights:

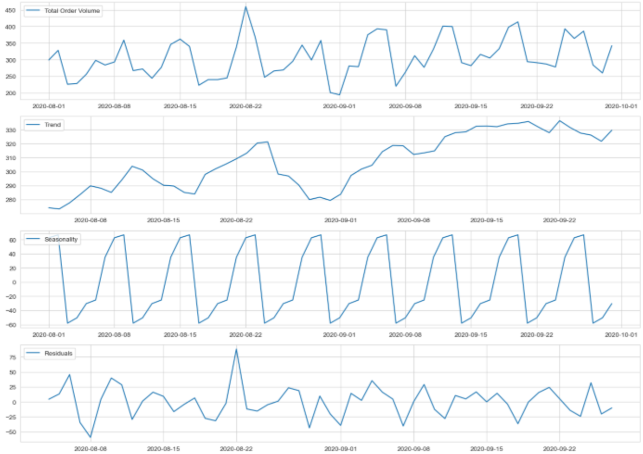
- **User Distribution Volume:** User density spans across bigger area of Helsinki (no clusters).
- **Venue Distribution Volume:** Clusters of high-order-volume venues.

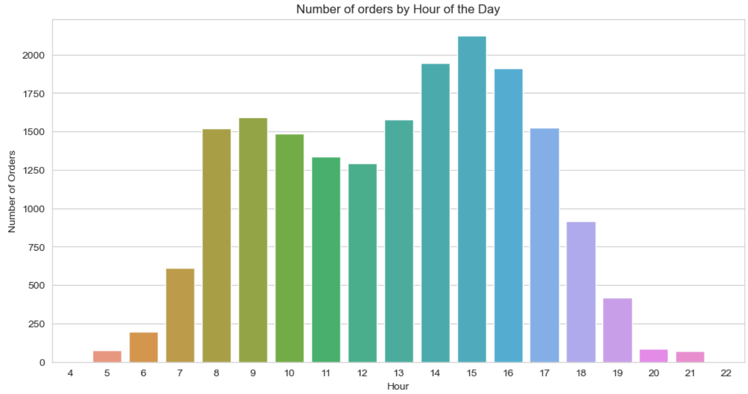


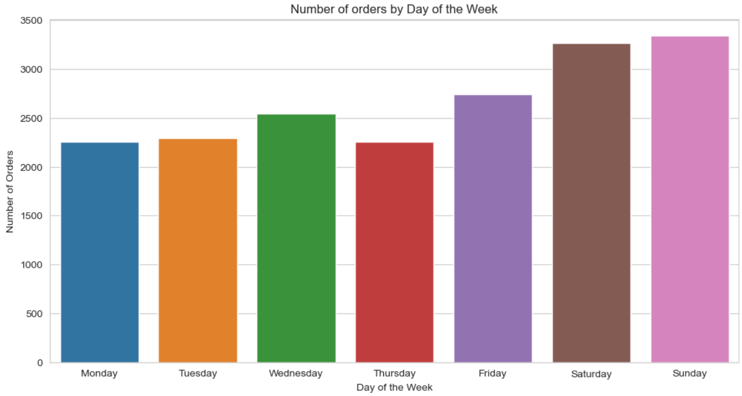


Temporal Patterns:

- **Peak Times:** 15:00 o' clock is the busiest time of the day. Weekends are the busiest days.
- **Seasonal Trends:** Increase in orders' trend (increase in September – colder months begin). Weekly trends are shown







Data Exploration (2)

Environmental Influence:

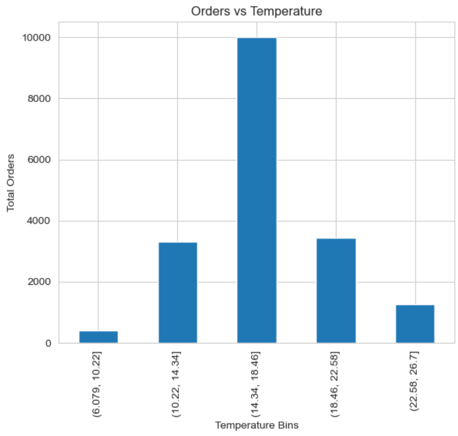
- **Weather Impact:** Not much influence on order volume.

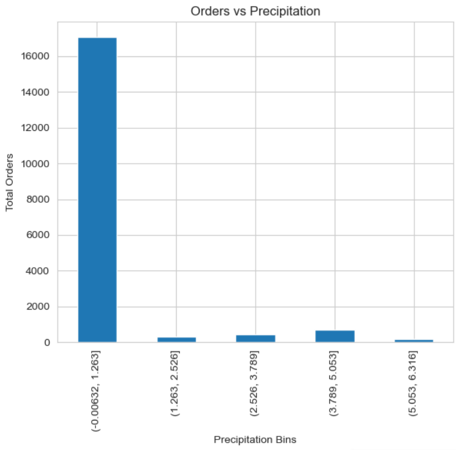
Feature Findings:

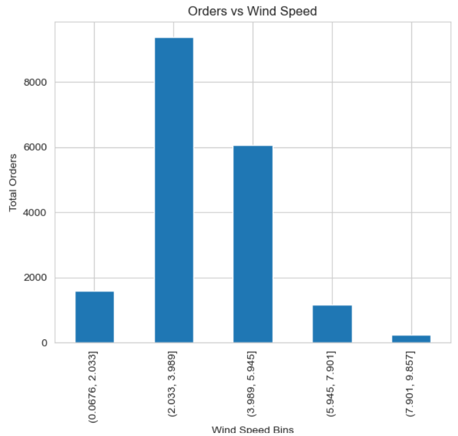
- **Correlation Insights:**
  - Delivery time features have a strong correlation with each other.
  - User and Venue coordinates have strong relationship.
  - Low correlation between weather conditions.
- **Feature Importance Analysis for Venue Popularity Prediction:**
  - Aggregated total orders by venue location.
  - Created target popularity tiers.
  - Introduced average user distance to capture user-venue proximity.
  - Applied K-means to form geographic clustering features.
  - Used Random Forest Classifier to understand feature importance. Also used permutation importance to assess the impact of each feature on predicting venue popularity.

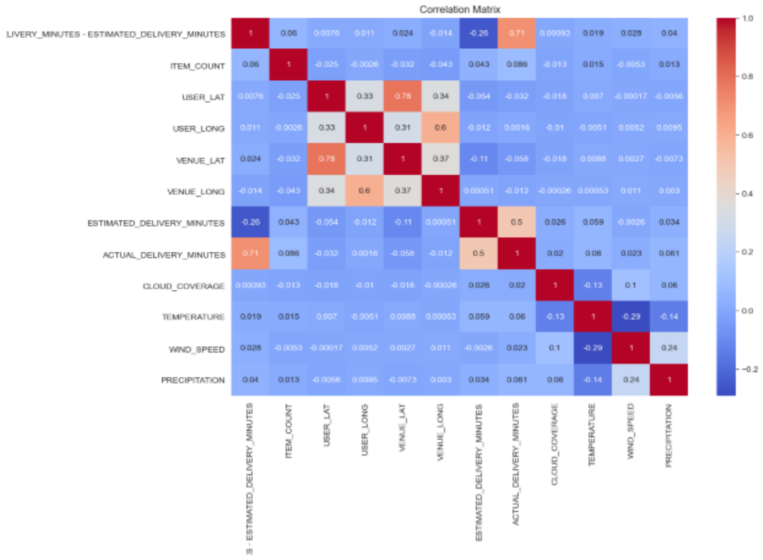
Discrepancies:

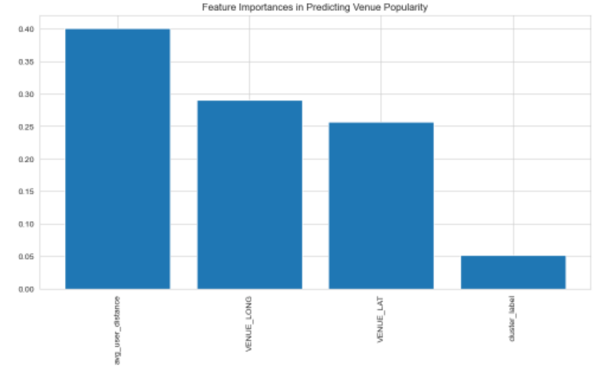
- **Outliers:** No outliers found.
- **Duplicated data:** No duplicated rows found.
- **Missing data:** 277 missing entries found for cloud coverage, temperature, and wind speed.

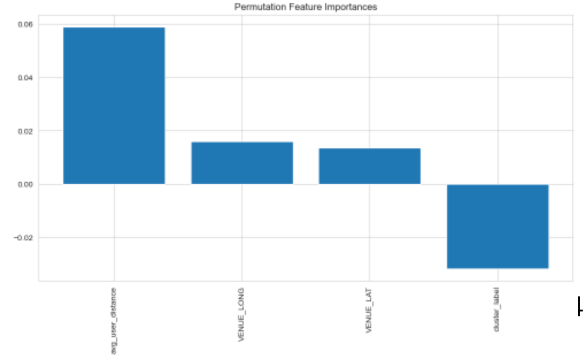












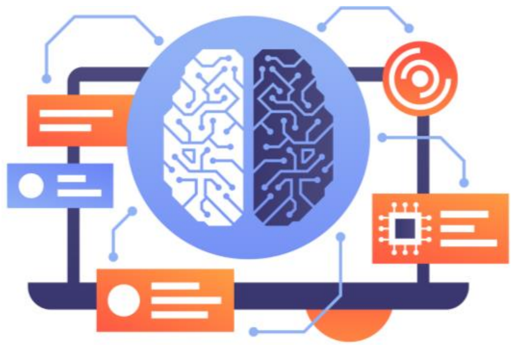
Feature Engineering and Modeling Approach

Feature Engineering

- **Geospatial Aggregation:** Orders were aggregated by venue location.
- **Popularity Tiers:** Venues categorized into 'Low', 'Medium', or 'High' popularity based on quantiles of order volume.
- **User Proximity:** Included the average distance between users and venues to capture potential influence on popularity. Utilized in further development.
- **Clustering:** Implemented K-means clustering to identify geographic hotspots, enhancing the model's spatial awareness. Utilized in further development.
- **Synthetic Venue Analysis:** Generated synthetic venue data to test model's predictive power on new locations, assessing its generalization capabilities and practical applicability.

Modeling Approach Rationale:

- **Predictive Task Relevance:** Features were chosen based on their expected influence on a venue's popularity, an important factor for operational and strategic decisions at Wolt.
- **Complexity Balance:** The approach strikes a balance between model complexity and interpretability, ensuring actionable insights.
- **Data Driven:** Clustering complements raw geospatial data, providing the model with structured spatial patterns that may not be immediately evident.



Models Used:

- **Naive Bayes:** Good and simple model with probabilistic insights (used as baseline).
- **Multilayer Perceptron (MLP):** Captures non-linear relationships and complex patterns through neural network architecture.
- **Random Forest Classifier:** Robust to overfitting, good for capturing intricate structures in the data.
- **Support Vector Machine (SVM):** Effective in high-dimensional spaces, suitable for clear margin of separation.
- **Gradient Boosting:** Builds strong predictive models through the ensemble of weak learners, optimizing on loss functions.
- **Ensemble Model:** Combines the predictions of individual models, aiming to improve accuracy and reliability.

Model Evaluation

Evaluation Metrics:

- Focused on Precision, Recall, and F1-Score from the Classification Report.
- These metrics were chosen to assess the balance between correctly identifying each popularity tier (precision) and the model's ability to detect all relevant cases of a tier (recall).

Naïve Bayes Results (Baseline):

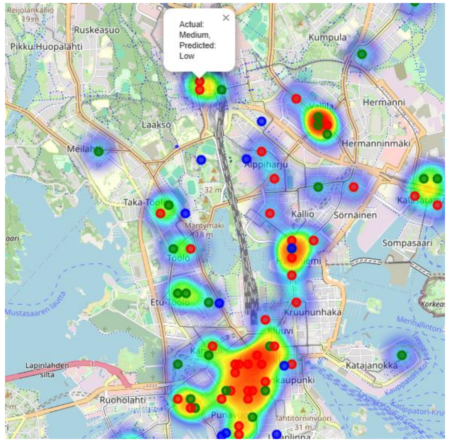
- **High Popularity Tier**: Precision - 0.32, Recall - 0.33, F1-Score - 0.33.
- **Low Popularity Tier**: Precision - 0.56, Recall - 0.50, F1-Score - 0.53.
- **Medium Popularity Tier**: Precision - 0.45, Recall - 0.46, F1-Score - 0.46.
- Overall Accuracy: 43%.

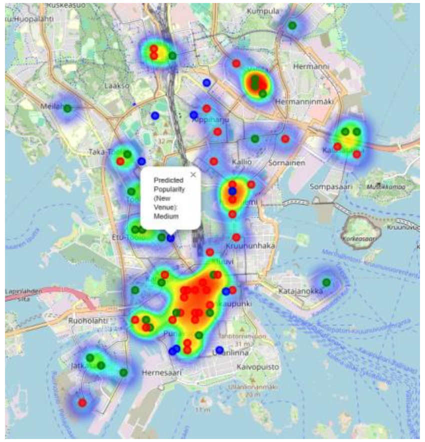
Insights:

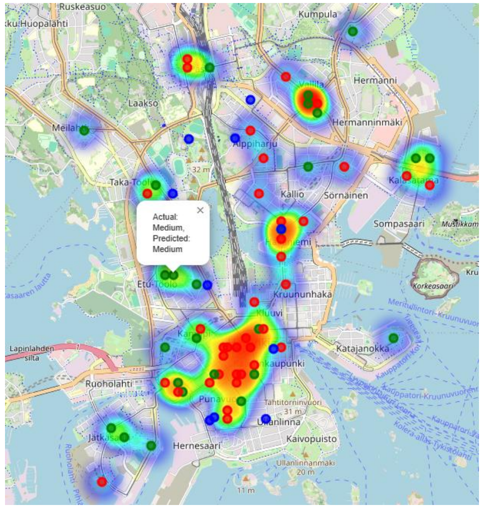
- The Naive Bayes model showed moderate performance with an overall accuracy of 43%.
- It performed best in identifying low popularity venues but struggled more with high popularity venues.

Visual Evaluation:

- Heatmap visualization highlights areas where the model correctly and incorrectly classified venue popularity (test set). Green marks for correct predictions, red for wrong, and blue for synthetic data predictions.
- This provided additional context to the numerical evaluation, helping to identify geographical patterns in the model's performance.







Further Development and Comparative Analysis

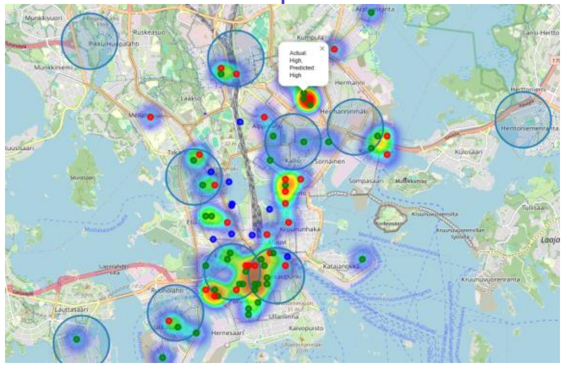
Feature Engineering Enhancements:

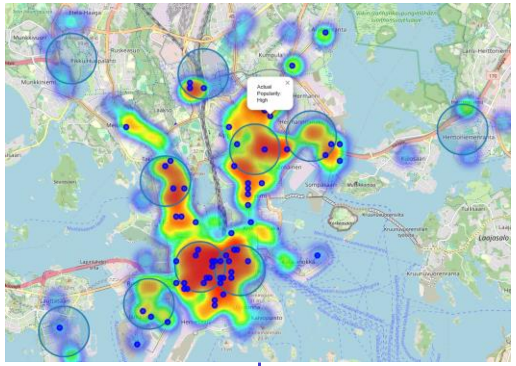
- Added **Geographic Clustering** and **Average User Distance** to provide spatial patterns and user-venue proximity.

Advanced Models Results using Macro Averages and 10 Clusters:

- **Ensemble**: Precision - 0.54, Recall - 0.54, F1-Score - 0.53, Accuracy - 54%.
- **Random Forest**: Precision - 0.54, Recall - 0.55, F1-Score - 0.54, Accuracy - 54%.
- **MLP**: Precision - 0.60, Recall - 0.56, F1-Score - 0.55, Accuracy - **58%**.
- **SVM**: Precision - 0.61, Recall - 0.52, F1-Score - 0.53, Accuracy - 51%.
- **Gradient Boosting**: Precision - 0.52, Recall - 0.53, F1-Score - 0.52, Accuracy - 52%.

Predicted venue popularity with clusters visible as well.





Actual venue popularity with blue marking the test set points. Points without markers are train set venues.

Insights:

- **Strengths:** Improved accuracy and balance in predicting popularity tiers.
- **Weaknesses:** Some models still struggled with certain tiers; complexity vs. accuracy trade-off.
- **Significant Difference:** Observed between Random Forest and SVM models (used p-value of 0.05).
- **Current Satisfaction:** The results show promising improvements over the baseline, particularly in the balanced accuracy achieved by the MLP model.
- **Production Expectation:** Provides valuable insights into venue popularity trends, aiding strategic decision making.

Future Development:

- **Data Enrichment:** Enhance model accuracy by incorporating additional data like venue types, customer reviews, and seasonal trends.
- **Real-Time Adaptability:** Implement models capable of utilizing real-time data to dynamically predict venue popularity, adapting to changing conditions and trends.
- **Continuous Improvement:** In a production environment, regular model updates and continuous monitoring will be essential to adapt to changing patterns and maintain high predictive accuracy.

Background and Aspirations at Wolt

Personal Introduction:

- **Name:** Kyriakos Kyriakou. My **current role** is Data Management Engineer at Sievo, Helsinki. So far, I got exposure in data modelling, data analysis, and consulting in procurement industry. My **education** background is Master's in Artificial Intelligence from the University of Edinburgh.

Passion for Data Science:

- **Thesis:** Collaborated with Amazon on Sequential Recommender Systems, exploring long-term prediction based on user interactions.
- **Skills:** Knowledgeable about various machine learning techniques in domains like Recommender systems, Advanced deep learning, Reinforcement learning, Speech Recognition, Vision, Robotics. Tech stack related to ML: PyTorch, NLTK, Scikit learn, Azure ML studio.

Passion for Data Science:

- **Content & Personalization Domain:** Passionate about improving customer experience through personalized recommendations. Experience from thesis.
- **Geospatial Data:** Recently developed a keen interest in working with geospatial data, as evident from the predictive model for venue popularity. Excited about the potential applications in Wolt's dynamic, location-based environment.

Thesis Motivation and Relevance to Wolt:

- **Cost-Efficiency and User Engagement:** My thesis approach focuses on predicting long-term user behavior, reducing the need for frequent model updates. This can lead to cost savings and more accurate user engagement predictions.
- **Future Work Potential:** Eager to explore and apply these insights in Wolt's environment, enhancing the ability to anticipate and meet customer needs effectively.

My Ambitions at Wolt:

- **Explore and Innovate:** Eager to contribute fresh ideas and approaches, particularly in personalized customer journeys and dynamic content adaptation.
- **Grow and Collaborate:** Looking forward to growing alongside Wolt's talented team, leveraging my skills in data science to solve exciting, real-world challenges.

Thank you