Are You Tired of Waiting for Your Food Order?

- *Imagine this real-world scenario:*
- After working tirelessly from morning to evening, you finally come home, exhausted and hungry. Cooking feels impossible after such a hectic day. You decide to order food using your favorite mobile app, only to find out that the delivery time is unusually long. The wait makes you even hungrier, angrier, and much more annoyed.
- This frustration may lead to customers choosing not to use the same food delivery application again.
- **But what if we could solve this problem?**
- Let's explore how we can tackle such issues and improve the customer experience with data-driven solutions!

Introduction

This report focuses on solving real-world problems in food delivery services by leveraging data science and machine learning techniques. Through rigorous data analysis and predictive modeling, aims to provide actionable insights to optimize delivery operations and improve customer satisfaction.

- **Steps Involved:**

1. **Dataset Preparation:** Cleaning and organizing raw data to ensure consistency and usability.
2. **Exploratory Data Analysis:** Uncovering patterns and trends in the dataset through visualization and statistics.
3. **Feature Selection and Statistical Analysis:** Identifying key features that significantly impact predictions and outcomes.
4. **Model Training and Optimization:** Developing predictive models and fine-tuning them for accuracy.
5. **Predictions, Questions, Results, and Insights:** Grouping data to understand delivery patterns and improve decision-making.
6. **Summary:** Drawing final conclusions and outlining the next steps based on insights and results.

Dataset Preparation

- **Dataset Selection:** Started with the dataset containing courier partner activity (daily_cp_activity_dataset.csv).
- The dataset was chosen for its focus on courier partner activity and relevant features like weather and day-of-week data, which directly impact delivery efficiency and customer satisfaction.
- **Data Cleaning:**
- Removed missing values and handled outliers.
- Generated cleaned datasets like **cleaned_daily_cp_activities.csv** and **cleaned_daily_cp_activity_no_outliers.csv**.
- **Feature Engineering:**
- Derived new features like **temperature_category** and **rain_intensity_category**.
- Prepared the **engineered_daily_cp_activity.csv** dataset.

Exploratory Data Analysis

Figure 1: Courier Activity Distribution
The distribution plot shows most of the courier activity is concentrated around specific levels, with occasional spikes. This indicates consistent demand with a few outliers that may represent unusually high or low activity days.

Figure 2: Seasonal Trends
The bar chart highlights variations in courier activity across seasons, showing higher activity during Spring and Summer, likely due to favorable weather conditions and increased customer demand.

Figure 3: Day-of-Week Analysis
A weekday activity analysis reveals higher courier availability on weekends, reflecting increased demand for food delivery during leisure days.

Figure 4: Precipitation vs. Courier Activity
A scatter plot between precipitation and courier activity shows a slight decline in activity during heavy rainfall, indicating adverse weather impacts on availability.

Figure 5: Temperature Impact
Courier activity peaks within a moderate temperature range. Extreme cold or heat negatively impacts courier availability, as shown in this scatter plot.

Figure 6: Correlation Heatmap
The heatmap shows a strong correlation between temperature, relative humidity, and courier activity, helping identify key features for modeling.

Figure 7: Box Plot for Outlier Detection
Box plots identify outliers in courier activity, which were cleaned during preprocessing to ensure data reliability.

Feature Selection & Statistical Data Insights

Evaluated feature importance using statistical methods and models. Key features like temperature, precipitation, and day of the week were selected for their strong correlation with courier activity.

- **Statistical Data Insights:**
  - **Correlation Analysis:** Temperature showed a positive correlation with courier availability, while precipitation had a slight negative impact.
  - **Feature Importances:** Random Forest highlighted temperature (48.46%) and relative humidity (15.42%) as the most impactful features.
  - **Outliers:** Box plots identified extreme values, which were addressed to improve data consistency.
  - **Seasonal and Weekly Trends:** Higher activity was observed during moderate weather and on weekends, reflecting customer demand patterns.
- **Why This Approach:**
  - Focusing on impactful features improves the model's interpretability and performance. Statistical insights help ensure data quality and identify potential modeling challenges.
- **Benefits:**
  - Simplifies the dataset, reducing computational load.
  - Enhances model accuracy by focusing on relevant predictors.
- **Evaluation Metrics:**
  - **RMSE (Root Mean Square Error):** Measures prediction accuracy.
  - **MAE (Mean Absolute Error):** Indicates average error magnitude.
  - **R\u00b2 Score:** Assesses the proportion of variance explained by the model.

Model Training and Optimization

Model Training:

1. **Linear Regression:** Trained as a baseline model due to its simplicity and interpretability. Achieved moderate accuracy with an RMSE of **7.58** and an R^2 score of **0.37**.
2. **Random Forest Regressor:** Leveraged for its ability to handle complex feature interactions. Results were less accurate with an RMSE of **8.37** and an R^2 score of **0.24**.
3. **Support Vector Machine (SVM):** Implemented to capture nonlinear patterns, but its performance was suboptimal, with an RMSE of **9.12** and an R^2 score of **0.09**.

Model Optimization:

1. **Random Forest Optimization:** Grid search was applied to fine-tune hyperparameters like the number of estimators and maximum depth. The optimized Random Forest showed minimal improvement, with an RMSE of **8.33** and an R^2 score of **0.24**.
2. **Feature Engineering Impact:** Encoding categorical variables and scaling numerical data improved model training and validation.

Key Insights:

1. **Linear Regression** (best_model.pkl) emerged as the most effective model due to its balance between accuracy and simplicity.
2. Optimization efforts highlighted the importance of feature selection and preprocessing in improving model reliability.

Questions, Results, and Insights

- **1. How many orders will we get tomorrow? Or next week?**
- **Results:** Predicted **72.03 orders** for the next day and weekly predictions of **67.39, 72.30, 74.79, 66.04, 65.93, 71.22, and 64.81 orders.**
- **Insights:**
  - Accurate short-term demand forecasts help allocate courier partners effectively, ensuring timely deliveries and reducing idle resources.
  - Weekly predictions enable planning for peak and low-demand periods, optimizing workforce management and operational efficiency.
- **2. Where will the orders be delivered in an hour?**
- **Results:** Cluster analysis identified delivery hotspots based on weather patterns.
  - Example: Areas experiencing high precipitation and low temperature (Cluster 1) are more likely to see increased delivery demand.
- **Insights:**
  - Enables better distribution of courier partners to high-demand areas, reducing delivery times.
  - Provides data-driven decision-making for expanding or adjusting delivery zones.
- **3. Based on past data, can we forecast the number of courier partners online for the next day, week, or longer?**
- **Results:** Daily and weekly courier availability forecasts were generated using Linear Regression.
  - Predicted availability for the next day matched closely with historical data patterns.
- **Insights:**
  - Accurate workforce projections help ensure adequate coverage during peak hours, improving

customer satisfaction.

- Prevents overstaffing or understaffing, reducing operational costs and enhancing resource allocation.

4. What is the expected revenue for tomorrow?

- **Results:** Predicted revenue for the next day is **$6346.91** based on predicted courier availability and order counts.
- **Insights:**
  - Provides reliable revenue forecasts for financial planning and budget allocation.
  - Identifies how external factors like weather and courier availability influence revenue, allowing for proactive adjustments to pricing and promotions.

Summary:

This assignment focused on solving the challenge of long delivery times in food services by applying predictive modeling and data analysis to streamline logistics operations. Accurate forecasts for order demand, courier availability, delivery locations, and revenue highlighted ways to reduce delivery times and enhance customer satisfaction. The results showed practical value, enabling efficient resource allocation and improved customer experiences if deployed to production. Future improvements could include advanced models like Neural Networks for better accuracy, real-time data integration for dynamic predictions, and external factors like traffic data. While the models are simple and interpretable, they face challenges with sudden real-world disruptions, leaving room for further refinement and automated decision-making systems.