MACHINE LEARNING ASSIGNMENT 1

Soluyanova Zlata Mikhailovna
z.soluyanova@innopolis.university

1 Motivation

Food delivery automation is becoming a crucial factor for enhancing service quality and optimizing business processes. The implementation of software and equipment allows for quicker order processing, minimizes errors, and improves customer interaction. This is especially important in a highly competitive environment where speed and accuracy of delivery play a decisive role.

Modern automation systems help manage couriers, optimize routes, and monitor order fulfillment in real-time. This not only improves service quality but also contributes to increased profitability, enabling restaurants to adapt to market demands and enhance their competitiveness.

2 Data

Regression:

In this task, the objective was to predict the actual preparation time (in minutes) using the following features: store_id, order_id, products, order_price, profit, delivery_distance, date_create, order_start_prepare, planned_prep_time, order_ready, order_pickup, region_id, and status_id.

Target: actual_prep_time

The analysis of correlation with the target variable revealed that the most influential features were:

planned_prep_time: correlation 0.623438
order_price: correlation 0.282730
products: correlation 0.237699
profit: correlation 0.234406
delivery_distance: correlation 0.119572

Features like order_ready, order_pickup, and others showed low or negative correlation with actual preparation time.

Classification:

In this task, the objective was to classify whether an order will be prepared on time or late.

Target: on_time

The analysis of correlation with the target variable revealed that the most influential features were:

Region ID: correlation 0.046471
Store ID: correlation 0.012745
Is Morning: correlation 0.007642
Planned Prep Minutes: correlation -0.132917
Delivery Distance: correlation -0.025121
Products: correlation -0.015353

Most of the signs have a weak impact on the timeliness of order preparation.

3 Exploratory data analysis

After removing outliers, the dataset was reduced to 148,490 entries and 24 columns. The average preparation time was 26.44 minutes in region 685 and 21.91 minutes in region 683, which also has the highest number of entries (191,240).

Missing values were identified, necessitating imputation or removal strategies. Outliers were detected using the Interquartile Range (IQR) method. Some features, such as delivery and profit, had their data types modified for consistency.

String representations of dates were converted to datetime format using pd.to_datetime(), after which the logical sequence of timestamps was validated. The columns status_id and order_id were removed: the former due to high redundancy and the latter because unique identifiers do not influence predictive modeling.

4 Task

Regression

I developed a regression model to predict Actual Preparation Time using features like store_id, order size, total price, and temporal data. The process involved loading data from SQLite, removing outliers with the IQR method, and filtering records. After splitting the data into training and testing sets, I imputed missing values (median for numerical, mode for temporal), scaled numerical features with RobustScaler, and encoded categorical variables using binary encoding. I evaluated Linear Regression, Ridge Regression, and Random Forest Regression based on MAE, MSE, and R^2. The selected model optimized operations and enhanced customer satisfaction.

Classification

I created new columns such as planned_prep_minutes and on_time, calculated the interaction between profit and delivery distance, and removed unnecessary columns. Outlier handling was done using the IQR method. The data was split into features (X) and target variable (y), followed by training/testing set division. Missing values were imputed (median for numerical, mode for datetime), numerical features were scaled with StandardScaler, and categorical variables were encoded with binary encoding. SMOTE was applied to balance classes. Models like logistic regression, random forest, and gradient boosting were trained and evaluated based on accuracy, precision, recall, F1 score, AUC-ROC, and confusion matrices. Cross-validation assessed model stability, and a bar plot compared mean F1 scores across models while identifying important features.

.

5 Input Format

I created a function, input_missing_values, to handle missing data more effectively than simple methods like mean or median imputation. My approach uses the median for numeric columns and the mode for datetime columns, which enhances robustness by minimizing the impact of outliers.This method also better preserves the original dataset's distribution, reducing bias that can arise from non-random missingness. It offers flexibility for incorporating advanced imputation techniques and retains all data points, maintaining sample size and statistical power—crucial for smaller datasets. My method provides more accurate and reliable results by considering data characteristics, reducing bias, preserving distribution, and maintaining sample size.

6 Comparison of selected ML models

Regression:

In my analysis of regression models using cross-validation, I evaluated Linear Regression, Ridge Regression, and Random Forest Regression, each showcasing unique strengths. The Random Forest model emerged as the best performer, achieving

the lowest mean cross-validation MSE, indicating its ability to capture complex relationships effectively. However, its high training R^2 raises concerns about potential overfitting.
Linear and Ridge Regression also performed well, with MSE values reflecting their effectiveness in capturing essential trends. Their simpler structures enhance interpretability, making them valuable for understanding model decisions.
To maintain robustness and avoid overfitting, I implemented strategies like regularization for Linear and Ridge Regression and hyperparameter tuning for Random Forest. I also explored data augmentation to enrich the training dataset and considered ensemble methods to combine model strengths.
I documented my findings with graphs and tables, including a boxplot showing MSE distribution across cross-validation folds and a table comparing training R^2 with mean cross-validation MSE. This analysis confirmed each model's capabilities and provided insights into their performance dynamics. The use of cross-validation and feature importance evaluation has equipped me with valuable knowledge for future modeling tasks.

Model	Mean Cross-Validation MSE	Training R^2
Linear Regression	23.3655	0.6869
Ridge Regression	23.3677	0.6869
Random Forest Regression	23.0225	0.9570

Classification:
Logistic Regression is the best-performing model with a mean F1 score of 0.5799 and a low standard deviation of 0.0035, indicating good generalization. Random Forest follows with a mean F1 score of 0.5570 but shows some instability due to a higher standard deviation of 0.0922, suggesting potential overfitting. Gradient Boosting has the lowest performance at 0.4955, likely indicating underfitting. To avoid overfitting and underfitting, strategies such as cross-validation, model complexity management, and regularization were employed. Overall, Logistic Regression stands out as the most reliable model based on cross-validation results.

Model	Mean F1 Score	Standard Deviation
Logistic Regression	0.5799	0,0035
Random Forest	0.5570	0,0922
Gradient Boosting	0.4955	0,0320

7 Results interpretation
Regression:
Significant features from Linear Regression, Ridge Regression, and Random Forest models reveal key factors influencing preparation time. Planned preparation time is a crucial predictor across all models, especially in Linear and Ridge Regression. Temporal features like prepare_month, create_month, prepare_day, and specific times of day (prepare_hour, create_hour) also play important roles.
Geographical factors include delivery distance, which appears in all models, indicating that longer distances may require more resources. Store ID and region ID suggest variations in operational efficiency. Days since month start is significant in both Linear and Ridge Regression, while order price and number of products influence preparation speed.
Ridge Regression emphasizes planned preparation time and delivery distance, whereas Random Forest captures complex interactions but highlights fewer features. The varying importance scores across models underscore the value of using multiple approaches to understand feature significance.
To optimize preparation processes, organizations should focus on planned preparation times and seasonal trends, while further analysis could explore feature interactions using advanced modeling techniques.

	MAE	MSE	R^2
Linear Regression	3.97	23.04	0.69
Ridge Regression	3.97	23.05	0.69
Random Forest Regression	3.93	22.80	0.69

Classification:
The analysis of influential features across different models reveals key insights into operational efficiency. For Gradient Boosting, planned_prep_minutes was the most significant predictor, followed by products and profit_to_price_ratio, highlighting the critical role of preparation time and product type. In Random Forest, delivery_distance and its interaction with profit were essential, indicating the importance of logistics, along with order_price. Lastly, in Logistic Regression, create_hour had a strong positive impact, while prepare_hour and planned_prep_minutes negatively affected outcomes, emphasizing that timing is crucial for success. Overall, optimizing preparation time, delivery distance, and order characteristics can enhance operational efficiency and customer satisfaction.

	Accuracy	Precision	Recall	F1 Score	AUC-ROC
Logistic Regression	0.5679	0.5305	0.5980	0.5623	0.5962
Random Forest	0.5756	0.5545	0.4339	0.4869	0.5956
Gradient Boosting	0.5792	0.5614	0.4259	0.4844	0.5963