Model Development Phase Template

Date	19 July 2024
Team ID	SWTID1720086535
Project Title	Ecommerce Shipping Prediction Using Machine Learning
Maximum Marks	6 Marks

Model Selection Report

In the forthcoming Model Selection Report, various models will be outlined, detailing their descriptions, hyperparameters, and performance metrics, including Accuracy or F1 Score. This comprehensive report will provide insights into the chosen models and their effectiveness.

Model Selection Report:

Model	Description	Hyperparameters	Performance Metric (e.g., Accuracy, F1 Score)
logistic regression	Uses a linear equation to estimate the probability of an order reaching on time based on the features which have coefficients indicating their influence on on-time delivery, good starting point for interpretability, but might not capture complex relationships between features.	-	Accuracy = 64%
logistic regression CV	An extension of Logistic Regression, model is trained on multiple subsets of the data, and its performance is evaluated on the remaining unseen data (cross-validation),	-	Accuracy = 64%

	helps prevent overfitting and improves the model's generalizability to unseen data.		
XGBoost	Builds multiple decision trees sequentially, where each tree focuses on improving the errors of the previous one, can handle complex non-linear relationships between features and can be very accurate but might be less interpretable than Logistic Regression	-	Accuracy = 66%
ridge classifier	Uses a linear equation but applies a penalty term (regularization) to control model complexity, helps prevent overfitting by reducing the influence of potentially irrelevant features, useful for datasets with many features or those prone to overfitting	-	Accuracy = 65%
K nearest neighbors	Classifies new data points based on the similarity of their features to existing labeled data points (on-time or delayed), simple to understand but can be computationally expensive for large datasets and sensitive to irrelevant features.	-	Accuracy = 63%
random forest	Ensemble learning method that builds a collection of random decision trees which predicts the delivery outcome based on a random subset of features, final prediction is the majority vote of all the trees, model offers good accuracy and handles non-linear relationships but can be less	-	Accuracy = 66%

	interpretable than simpler models.		
support vector classifier	Finds a hyperplane that best separates the data points representing on-time and delayed deliveries based on their features, works well with high-dimensional data but can be sensitive to feature scaling and might be computationally expensive for large datasets.	-	Accuracy = 66%