Checkpoint 2

BFI-Sustainability Team

Team Members - Akash Kesani, Ravindu Tharanga Perera, Samuel Kannan

Justification of Model Choice

We chose Logistic Regression as our model for the data for the following reasons:

1. Nature of Target Variable: Our response variable "Status" represents a categorical outcome (multi-class), Logistic Regression is inherently designed for such scenarios. It models the probability of the target variable categories as a function of the independent variables, making it a natural fit for predicting categorical outcomes.
2. Relationship between variables: Logistic Regression can effectively capture linear relationships between the log odds of the outcomes and the predictor variables. If the relationship between outcome variable "Status" and the predictors like "Priority", "Subcategory ID", and "Organization ID" approximates a linear relationship in the log-odds space, Logistic Regression will perform well.
3. Simplicity and Interpretability: Logistic Regression is relatively straightforward to implement and interpret. It provides coefficients that represent the log odds impact of each predictor variable, making it easier to understand the influence of each feature on the probability of the outcomes. This interpretability is particularly valuable in many practical applications where understanding the model's decision-making process is as important as the model's accuracy

Model's Test Error Rate

We are calculating the accuracy as a measure of test error rate since it is a classification task. We use the accuracy_score(y_test, y_pred) function from the scikit-learn library and computes the accuracy of the model's predictions by comparing the predicted labels (y_pred) to the actual labels (y_test).The resulting accuracy variable will contain the accuracy score, which represents the proportion of correctly classified instances out of the total instances in the test set.For our model, we get accuracy score of 0.887

Model's Goodness of Fit

Our accuracy score of 0.887 suggests that the model fits the data well, especially considering the complexity and potential variability in the response variable "Status". It's a strong indication that the chosen predictors ("Priority", "Subcategory ID", "Organization ID") have a significant relationship with the target variable and that logistic regression is a suitable modeling approach for this problem. Further analysis could involve examining the confusion matrix, precision, recall, and F1 score to better understand the model's performance across different "Status" categories. Additionally, exploring other models or feature engineering could potentially improve the model's performance. However, the current model's accuracy suggests a strong starting point and indicates a meaningful relationship captured by the Logistic Regression model between the features and the target variable.

Cases of Interest

In our data we find three interesting problems that can be solved by machine learning models. We formulate the following three problems from our data and how our model solves the problem.

1. Large Dataset: We have 377,060 observations in the dataset making it quite large so we'll need to select a model that is fast to train and predict. Logistic regression overcomes this problem by being generally fast to train and predict, making them suitable for situations where computational efficiency is important.
2. Imbalanced classes: In our data, our outcome variable "Status" is an imbalanced class where one class is significantly more prevalent than the other(s). In our dataset Status of '3' is more prevalent than other classes. Logistic regression is helpful in dealing with imbalanced classes using techniques like class weights. In our data, we assigned different weights to each class during model training. By assigning higher weights to the minority classes like '1' and '2', logistic regression focused more on correctly classifying instances from the minority classes, thus mitigating the impact of class imbalance.
3. Flexibility: Till now, the relationships between predictors and outcome variables have been a linear relationship. In the future, if we are interested in predictors that will have a non-linear relationship with the outcome variable, then the model must be flexible to accommodate that. Logistic regression solves this problem, as it can be extended to handle non-linear relationships through feature engineering, such as polynomial features or interaction terms. This allows for flexibility in modeling more complex relationships without moving to more computationally intensive models