Cloud Final Project Write Up Question 1

Kevin Nguyen, Abhigyan Acherjee

The following’s a short overview of three popular machine learning models: linear regression, random forest, and gradient boosting algorithms:

i. Linear Regression

Linear regression is one of the simplest and most widely used statistical techniques for predictive modeling. It attempts to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. Linear regression is best used for scenarios where data exhibits a linear relationship, and it's often the first algorithm applied in predictive modeling tasks.

ii. Random Forest

Random forest is an ensemble learning method primarily used for classification and regression that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random forests correct for decision trees' habit of overfitting to their training set. Each tree in a random forest is built from a sample drawn with replacement (i.e., a bootstrap sample) from the training set. Moreover, when splitting a node during the construction of the tree, the split that is chosen is no longer the best split among all features; instead, it is the best split among a random subset of the features. This results in a forest of trees that are not correlated, enhancing the predictive accuracy and control over-fitting.

iii. Gradient Boosting Algorithms

Gradient boosting is a powerful ensemble machine learning technique that builds models incrementally in a stage-wise fashion. It involves three elements: a loss function to be optimized, a weak learner to make predictions, and an additive model to add weak learners to minimize the loss function. Each new model takes a step in the direction that minimizes the prediction error, as indicated by the gradient of the loss function. Popular variants of gradient boosting include XGBoost, LightGBM, and CatBoost, which have proved effective in a wide range of practical applications. Gradient boosting can be used for both regression and

classification problems, providing a highly flexible framework that can handle various types of data and distributions.

These models cover a spectrum from simple to complex and are used extensively across different fields for numerous applications, from predicting housing prices to classifying objects in images. Each has its strengths and ideal use cases, making them staples in the toolbox of data scientists and machine learning practitioners.

For this project, I would use a combination of linear regression and knn to silo the sales into different categories.