Final Project ML Write up

Cloud Computing Spring 2025

Group 1 – Chloe Belletti, Preston Buterbaugh, Aakash Rammohan

Linear Regression:

Linear Regression is a simple and widely used method for predicting a continuous outcome by drawing a straight line that best fits a dataset. It finds the line's intercept and slope by using techniques like ordinary least squares and gradient descent, to find the best-fit line $y = B_0 + B_1*x$ [1]. This allows each coefficient to clearly show how much its input affects the result. It's often the first choice for regression problems because of its ease of use and speed. To improve the model's accuracy, it's important to remove redundant information and clean out noisy data.

Random Forest:

Random Forest is an ensemble technique that makes predictions by averaging the results of many decision trees. It creates each tree from a random sample of given data and a random subset of features, then it combines their outputs using averages for regression or a majority vote for classification. This randomness helps reduce errors and overfitting, while still capturing complex patterns. Random Forest also provides feature-importance scores so you can see which inputs matter most. It's often chosen for its ease of use.

Gradient Boosting:

Gradient Boosting builds a strong predictor by adding many simple models one after another, where each new model focuses on fixing the errors of all the models so far. It starts with a basic learner, usually a small decision tree and then trains each subsequent tree on the remaining errors. All tree predictions are summed, and a learning rate shrinks each tree's impact to help prevent overfitting [1]. Libraries like XGBoost, and CatBoost enhance this process with regularization and parallel processing. Gradient Boosting can uncover complex patterns and feature interactions, but it needs clean data and careful management of its parameters to work best.

Customer lifetime value analysis:

Random Forest regression is ideal for predicting Customer Lifetime Value because it captures complex, nonlinear relationships among features such as purchase frequency, recency and the monetary value of items. By averaging many decision trees trained on bootstrap samples and random feature subsets, it reduces variance and handles outliers and missing data well. Our team used the Random forest technique to train a CLV model as it works best with the data we were given.

References

[1] “Top 10 machine learning algorithms to know,” Built In, https://builtin.com/data-science/tour-top-10-algorithms-machine-learning-newbies (accessed Apr. 25, 2025).