Project Proposal

Proposed Topic:

Boosting Ensemble Learning: A case study of "XGBoost" for Supermarket Sales Prediction

Motivation

This project intends to examine and evaluate the performance of XGBoost, an ensemble machine learning algorithm based on boosting, on a supermarket sales dataset. Boosting is an ensemble learning technique that seeks to minimize training errors by building a powerful classifier out of a number of weak classifiers [1], [2], [8]. Firstly, a model is built on the training dataset, then another model is built on top of that to try to correct the errors of the previous model, this procedure is repeated until either the maximum number of models are added or training is completed [3], [5]. Similarly, XGBoost which stands for eXtreme Gradient Boosting was specifically designed to improve speed and performance, it is now a popular and efficient open-source implementation of the gradient boosted trees algorithm.

Over the past few years, XGBoost has gained traction in the data science/data mining ecosystem by helping individuals and teams on Kaggle win almost every structured tabular data competition as well as by being actively utilized by multiple organizations such as Delivery Hero, Compile Inc, BagelCode, BlaBlaCar etc. [3]. Its opensource characteristics has also resulted in a rising number of data scientists globally that are actively contributing to improving the codebase. For this project, I intend to apply XGBoost algorithm to build a predictive model to find out the sales of each product at a supermarket. I would also evaluate the trained model using the Root Mean Squared Error (RMSE) value and compare it against a baseline linear regression model.

Problem Definition

I am interested in evaluating the performance of an ensemble technique, specifically the XGBoost model on a supermarket sales dataset, and comparing the results to a baseline regression model. For this project scope, the input data is a supermarket sales dataset that will be obtained from Kaggle. This data will first be fed to a linear regression model to get a baseline Root Mean Square Error (RMSE) value, and then to an XGBoost model for an even better score. The outputs will be a trained model, and the output predictions of trained model.

Methods

After getting a baseline RMSE score from a linear regression model, I intend to apply XGBoost on the dataset to examine and evaluate its performance based on the RMSE as the evaluation metric. XGBoost is integrated with sci-kit learn for Python enthusiasts, sci-kit learn is a very robust and efficient library for machine learning in Python. For this project I will also be using sci-kit learn, pandas, and Jupyter notebooks to create a prototype and compile the codes for predictive analytics and evaluation.

Evaluation Criteria

The evaluation criterion will be the Root Mean Squared Error (RMSE), which is the standard deviation of the residuals. The RMSE is a measure of how dispersed the residuals are, or how consolidated the data is around the line of best fit in a regression model [11].

References

[1] https://www.geeksforgeeks.org/boosting-in-machine-learning-boosting-and-adaboost/

[2] https://www.nvidia.com/en-us/glossary/data-science/xgboost/

[3] https://stackshare.io/xgboost

[4] https://www.ibm.com/cloud/learn/boosting

[5] https://www.mygreatlearning.com/blog/xgboost-algorithm/

[6] https://www.kaggle.com/competitions/dsn2018intercampus/overview/evaluation

[7] https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html

[8] https://www.nvidia.com/en-us/glossary/data-science/xgboost/

[10] https://stackshare.io/xgboost

[11] https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/

Victor Irekponor | Ph.D. student

119079201