Mentee: Liuyixin Shao

Mentor: Charlie Wolock

SPA DRP, winter 2023

Introduction to Prediction

This is a very meaningful quarter for me because I did learn a lot under mentoring of Charlie.

We first went through chapters 1, 2, 3, 4, 5, 8, and 10 in the book “An Introduction to Statistical Learning”. I really like the book because it teaches me a lot of advanced statistical concepts and their applications in the real world. Specifically, we learned concepts such as bias-variance tradeoff, linear regression, logistic regression, bootstrap, cross-validation, decision tree model, random forest tree model, boosting model, convolutional neural network, and recurrent neural networks. At the same time, Charlie's unique teaching style honed my ability to summarize knowledge from pages in the book and describe it concisely for others.

Then, I applied what I learned to a project to predict Airbnb prices in New York City. In this project, I first one-hot-encoded my data, removed the outliers, and avoided correlation between all the variables I chose for the prediction. Then, I constructed four different models including linear regression, decision tree, random forest, and gradient boosting. After that, I modified each model with proper strategies such as removing outliers and high leverage points to build the linear regression model, applying 5-fold cross-validation to evaluate model performance and avoid overfitting and calculating out-of-bag error from bagging to test the model accuracy. In the end, my random forest model has the best result with the lowest testing mean squared error.

Before participating in the Direct Reading Program, as a statistics major who only took 2 college-level statistics classes, I felt very confused about my future because I had very little understanding about my major and didn’t know any application of statistics. However, after the guidance of Charlie, I realized the massive use of statistics in machine learning, neural networks, and AI and found my true interest-statistical learning. I appreciate the opportunity to learn about advanced-level statistics and I deeply thank my mentor Charlie Wolock for being so patient with me throughout the quarter, helping and supporting me until the end.

In the end, I listed some of the explanations of the concepts that I’ve learned throughout this quarter.

The Bias-Variance Trade-Of: As a general rule, as we use more flexible methods, the variance will increase and the bias will decrease. Here, bias refers to the error that is introduced by approximating a real-life problem. Variance refers to the error that is introduced by the model's sensitivity to the variability in the training dataset.

Linear Regression: Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on X_1, X_2, …, X_p is linear. The goal of linear regression is to find the coefficients that minimize the squared difference between the predicted values of Y and the actual values.

Logistic Regression: logistic regression models the probability that Y belongs to a particular category using a logistic function. The goal of logistic regression is to find the relationship between the input variables and the probability of the response variable being equal to 1.

Bootstrapping: Bootstrapping is the method to create multiple resamples of the original dataset by randomly sampling with replacement from the original dataset. In this case, it quantify the uncertainty associated with a given estimator or statistical learning method.

K-fold cross-validation: K-fold cross-validation involves dividing the data into k equal-sized subsets, or folds, and using each fold once as a validation set and the remaining k-1 folds as the training set. We use it to evaluate model performance and avoid overfitting

Decision Tree Model: The Decision Tree model is built by recursively splitting the data into smaller subsets based on the values of the input variables. Each split corresponds to a decision on the input variables, and the resulting subsets are used to further refine the decision tree. At the end of the tree, the model produces a prediction based on the values of the input variables. It's possible for the decision tree to be overfitted. Thus, we need to prune the tree using K-fold cross-validation.

Random Forest Model: The basic idea behind a random forest is to combine multiple decision trees into a single model that can make more accurate predictions than any individual tree. Based on the idea of bagging, each decision tree in the random forest is trained on a random subset of the data and a random subset of the input variables. This randomness helps to reduce overfitting and makes the model more robust to noise and outliers in the data. The final prediction of the random forest is based on the average prediction of all the trees in the forest, weighted by their individual performances.

Out-of-bag error: a biproduct of bagging. Because every time we use a subset of data to train the model, we can use the rest of the data as validation and test the model accuracy.

Boosting Model: In boosting, each tree is grown using information from pervious tree. The algorithm works by iteratively adding new trees to the model, with each tree attempting to correct the errors of the previous trees. Here, we also want to tune the parameters for boosting.

Convolutional Neural Network: Convolutional neural network is designed to recognize visual patterns directly from pixel images. By alternatively using convolutional layers and pooling layers in the neural network, it can extract different features from the input data and sharpen each feature's identification.

Recurrent Neural Network: Recurrent neural network is designed to handle sequential data. It builds models that take into account the sequential nature of the input data and builds a memory of the past. Each neuron in an RNN receives input from both the previous time step and the current time step, and produces an output that is used as input for the next time step.