**Tutku Kılıçaslan**

**54201**

**ENGR 421**

**Homework #6**

**DIVERSITY IS THE BEST WAY TO LEARN**

**Objective**

The main objective is to predict the amount that will be withdrawn from the ATMs in given region in a given period and transaction type. Our measurement will be mean absolute error (MAE) and root mean squared error (RMSE) of our predictions.

**Observations**

As I observed, there are 20 different region with 2 ATMs except region 2 (3 ATMs), region 13 (4 ATMs), region 19 (4 ATMs) and region 20 (4 ATMs). There are two different transaction type which are card present and card not present. We have also the transaction date as data.

**Brain storm**

* Most proper algorthms for multiple regression are **decision tree, MLP and regression.**
* We can get higher accuracy by using **different algortihms**.

Combination of different algorithms gives us higher accuracy. For our data decision tree and MLP can be suitable.

* We can get higher accuracy also by using **different parameters**.

Some algorithms are sensetive to parameters. Therefore combination of algorithm with different parameter can give us better results.

* We can get higher accuracy by using **different input representations.**

We can give different input to the algorithm and we can get higher accuracy by using these different results from different algorithms. For our data, we can use random forest method to use different input representation.

**My Solution Method**

Since our problem is predicting, I decided to devide my data into training and validation sets to evaluate if my model correctly predict the given data. Therefore, I constructed a large window and slide the window 10 days to predict next 10 days. By implementing this I sacrifised my data of 10 days. However, this sacrifice can be ignore since my training data is huge enough.

Before starting I ordered my data by date. By doing this, I had my data as a time series problem. After that, I choose the data before last 20 days in training set as training data as input for my random forest algorithm and used the next 10 days as validation data. I first kept the parameters as default but by using **ntree** and **mtry** parameters, the model can be improved even better. These improvements and parameter selection will be discussed later in **Outcomes** section.

As I implemented the random forest model, I predicted validation data and found the errors as **MAE=10.95065** and **RMSE= 18.32056.**

I slided my window 10 days more and implement the same procedure for the data after first 10 days. I implemented the same algorithm and obtain a new model.

After predicting the last ten days in my training data, I found even better results as follows.

**MAE= 10.94831** and **RMSE=18.30471.**

**Therefore, I assumed that the next 10 days can be assumed by this model. I used the data after first 20 days in my training data and estimate the transactions next 10 days by implementing random forest.**

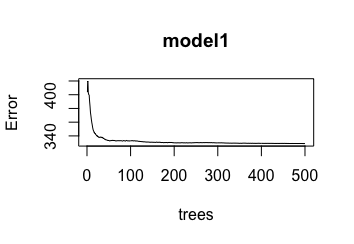
**Outcomes**

Random forest algorithm needs parameters **ntree** and **mtry**.

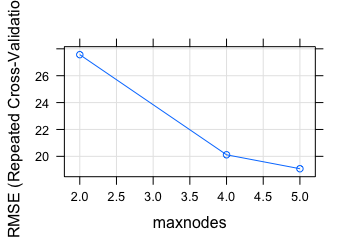
**mtry**: Number of variables randomly sampled as candidates at each split.

**ntree**: Number of trees to grow.

If we increase ntree this increase our diversity. However, we saw that after certain amount ntree doesn’t change our error that much as the graph below shows. Since the value drastically change the processing time, I choose **100** as ntree.

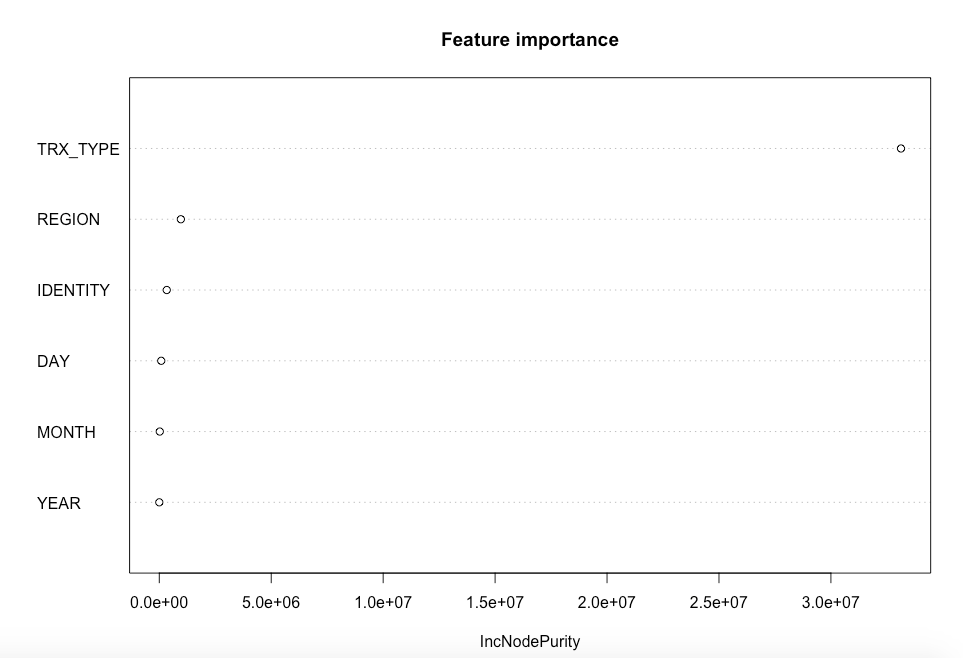


In terms of mtry, as we increase mtry we expect to learn better since the depth of our trees will be greater. As we saw on the graph below, for a constant ntree, higher mtry gives us higher accuracy. However, this also drastically change our processing time. Since our number of features 6 and 2-3 of them has visible change in our prediction, I choose mtry as 3 by considering both speed and accuracy. This is discussed more at the end of this section.



To implement larger ntrees and mtrys will result in longer process time. It is a trade off that prefering either accuracy or speed. In my solution, I tried to optimize both of them as much as I could. Therefore, my solution takes some time. If our data had been larger than this, this could have taken much more time.

By implementing grid search algorthim, I obtained the graph below. According to the graph, the transaction type plays the most significant role for future prediction compared to other features. The least significant feature is year since we have only one unique year value in our training data.



**Improvement**

In the solution key, we were given decision tree regression model. The solution was quick and dirty. Since our measurements are MAE and RMSE I tried to improve those by implementing random forest.

I improved the prediction MAE from ~14.465 to ~10.706 and RMSE from ~23.728 to ~18.270.

This means **~26%** improvement in MAE and **~23%** improvement in RMSE.

**Conclusion**

During the semester, we have learnt to implement different algorithm. However, one algorithm does not always give the best solution. We can improve our solution by taking the advantage of diversity. In my solution, I improved the diversity of my input for decision tree. Since the decision tree algorthm is a unstable algorithm, I performed better prediction by using random forest method. My measurement was **MAE** and **RMSE**. At the end of the implementation, I obtained **~26%** and **~23%** decrease respectively. It is important that this solution may not be the best solution but it solves our problem and improve the error with a good amount.

**References**

* <https://cran.r-project.org/web/packages/randomForest/index.html>
* <https://machinelearningmastery.com/tune-machine-learning-algorithms-in-r/?unapproved=515309&moderation-hash=ef84825e5d8bad58fa9487b5133a239f#comment-515309>
* <https://www.r-bloggers.com/how-to-implement-random-forests-in-r/>
* <https://uc-r.github.io/random_forests#idea>
* <https://www.datacareer.ch/blog/random-forest-in-r-an-example/>