Review of Random Forests technique and the Application of Random Forests in Travel Time prediction

Random Forests

In Adaptive Business Intelligence

ISCG 8043 Adaptive Business Intelligence

Chunkai Meng

Student ID: 1494933

Random Forest in Adaptive Business Intelligence

**Table of Contents**

[*Abstract-* 2](#_Toc510386876)

[1. Introduction 2](#_Toc510386877)

[1.1 Summary and background 2](#_Toc510386878)

[1.2 Why it is important 2](#_Toc510386879)

[1.3 How it has been studied or developed 2](#_Toc510386880)

[2. Concept and Theory 2](#_Toc510386881)

[2.1 Description of the Concept and Theory 2](#_Toc510386882)

[2.2 How the Concept and Theory related to ABI 3](#_Toc510386883)

[2.3 Advantages of the Random Forests technique 4](#_Toc510386884)

[2.4 Disadvantages of the technique 4](#_Toc510386885)

[2.5 Limitations of the technique 5](#_Toc510386886)

[3. Application: 5](#_Toc510386887)

[3.1 A specific example application related to Adaptive Business Intelligence 5](#_Toc510386888)

[3.2 How to apply the Random Forests on the Travel Time Prediction 5](#_Toc510386889)

[3.3 Advantages of Random Forests technique on Travel Time Prediction 6](#_Toc510386890)

[3.4 Disadvantages of Random Forests technique on Travel Time Prediction 7](#_Toc510386891)

[3.5 Other available techniques that can be used for the travel time prediction 8](#_Toc510386892)

[4. Future Development and Conclusion: 8](#_Toc510386893)

[4.1 The future of the technique 8](#_Toc510386894)

[4.2 The future of the technique in relate to ABI 8](#_Toc510386895)

[4.3 Conclusion: 9](#_Toc510386896)

# Abstract:

# 1. Introduction

## 1.1 Summary and background

We are living in the century of big data. The effective use of data is becoming the key competency of almost any business in any industry. To make the right choices at the right time from data, the concept of Adaptive Business Intelligence (ABI) is created. Apart from the original Business Intelligence (BI), ABI gives more attention to the facts that data are increasing and changing dramatically in every minute. This ABI system must be able to self-learn so that it can give real-time and near-optimal the solutions or decisions reflecting these changes.

## 1.2 Why it is important

Random Forests is the ensemble of classification or regression trees, which means they are a combination of relatively simple models. By increasing the number of the tree, the overfitting is effectively controlled or eliminated(Breiman, 2001) . In modern ABI Systems, Random Forests model is becoming more and more important due to its comparable accuracy, the wide range of uses and low computational in training and predicting.

## 1.3 How it has been studied or developed

In recent years, Random Forest has been studied and applied a lot in many areas. Kim and Giles (2016) proposed a new algorithm based on Random Forests to discover the relationship of Financial Record stored in different databases. Hu, Chen, Hu and Peng (2018) designed a novel random forests-based class incremental learning method to recognize human activity automatically. Zhang, Cao, Li and Wang (2018) proposed a Cascaded Random Forest method to classify hyperspectral images. Liu, Cao, Zhao, Mulligan and Ye (2018) applied Random Forests method into their geostatistical research, thus improved ground-level PM 2.5 concentration mapping accuracy.

# 2. Concept and Theory

## 2.1 Description of the Concept and Theory

Random Forests are usually referred as an ensemble of Decision Trees. The image below illustrates an informal Decision Tree which is a classification tree (Classifier).

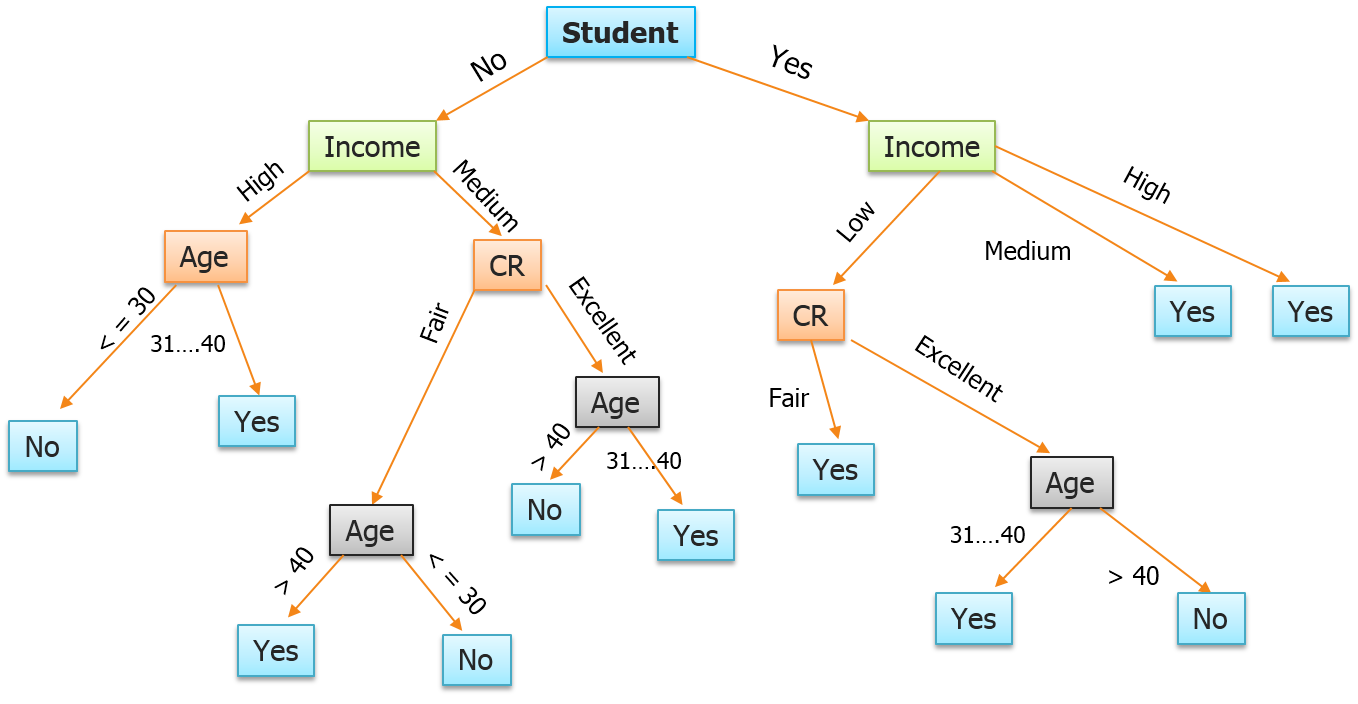


Figure 1: Decision Tree (Akansha, 2015)

The process of creating a Random Forests can be simplified as below:

1. Use bootstrap aggregating (or bagging) method to select the number K of new training datasets randomly.
2. In each training set selection, those samples have never been selected form the out-of-bag data set. Which then be used for validation.
3. In each tree construction, "m" features with most influences from total "n" features are used to create the decision tree.
4. Build the forests by repeating the step 3.
5. Calculate the votes of each tree
6. Use the high voted predicted result as final prediction.

For each node, features are select to split the data into classes and decided which class this node belongs to. For example, in the above Decision Trees, nodes are firstly divided into two classes by the feature that whether it is a student or not. Random Forests combine a set of small Decision Trees which created by embedding bootstrap random sampling and adopting Gini index as a feature selection standard (Hu et al., 2018) .

## 2.2 How the Concept and Theory related to ABI

In ABI systems, data mining plays a very important role, which generates knowledge from information. There are two kinds of data mining, verification and discovery regarding its mining goal. Verification data mining methods aim to the goodness of fitting hypothesis, while discover methods are used to identify patterns. These result patterns can help us either describe the underlying rules or predict the values we are interested. The discover method for predicting are also mentioned as supervised learning or supervised model. It can be divided into Classification Model and Regression Model, these are the most popular kinds of a model in Data Science. Decision Tree is the approach that can be adopted in both Classifications as well as Regression (CARD) (Rokach & Maimon, 2014). Apart from the traditional role of BI, ABI gives more emphasis on building decisions automatically and continuously. There are a large number of studies on Random Forests in this kind of areas. Thanks to many state-of-the-art algorithms, Hu et al. (2018) state that the Random Forests has become an efficient and effective learning method for online learning and incremental learning domains.

## 2.3 Advantages of the Random Forests technique

In the ABI domains, Random Forest has some key advantages as below:

1. Random Forests technique is suitable for both categorical data and continuous data.
2. High training speed: ABI systems are required to handle large data and or even data streams. Components of Random Forests are independent trees, they can be built simultaneously on different CPU cores. This dramatically increases the speed of decision generation (Genuer, Poggi, Tuleau-Malot, & Villa-Vialaneix, 2017).
3. High accuracy: Decision Tree method is simple and strict-forward, however, a slight change of the data can drastically affect the features selection and so affect the whole structure of a tree. By subsampling N decision trees, the potential variance and bias are reduced and removed (Breiman, 2001).
4. Random Forests can deal with complex data. High dimensional data, which means the number of features is relatively big (Breiman, 2001).

## 2.4 Disadvantages of the technique

Random Forests have several disadvantages such as:

1. Compared to the other model, people cannot acquire too much insight from the Random Forest output, some engineer believe it’s a black box that are hard to interpret.
2. Parallel training Random Forests maybe not applicable in a real scenario or constrained by its sample size (Genuer et al., 2017).
3. The memory consumption could be problem if the depth and or the number of tree is not optimal.

## 2.5 Limitations of the Random Forests

1. When the relationship between dependent and independent variables is highly linear, Random Forests could be relatively week as a regression model compared to classification model.
2. Random Forests technique is not suitable for small dataset learning, because it accuracy may decrease dramatically.

# 3. Application:

## 3.1 A specific example application related to Adaptive Business Intelligence

Adaptive Business Intelligence System should be able to deal with real problems in the time-changing world and should be resistible to the complexity of vast constrains. Prediction and Optimization are the keys to those techniques or models involved in Adaptive Business Intelligence (Michalewicz, Schmidt, Michalewicz, & Chiriac, 2006) . In modern cities, travel time is the most important criteria for the traffic condition. It usually reflects the city development level. In contrast to other intelligence systems that only demonstrate the current congestion or recurring rush hour on the roads. A real-time prediction on travel times has great significance not only to the drives but also to the traffic light systems and urban planning. In recent years, many studies focus on using Random Forest to predict the travel times.

## 3.2 How to apply the Random Forests on the Travel Time Prediction

Hou, Edara, & Chang (2017) studied the Road Network State Estimation using Random Forest Ensemble Learning. The traffic data came from the St. Louis, USA. They are real traffic data in two years, from June 2014 to May 2016. Firstly, the researchers pick up one load as the predicting target, which is F3 shown in red colour in the image below:

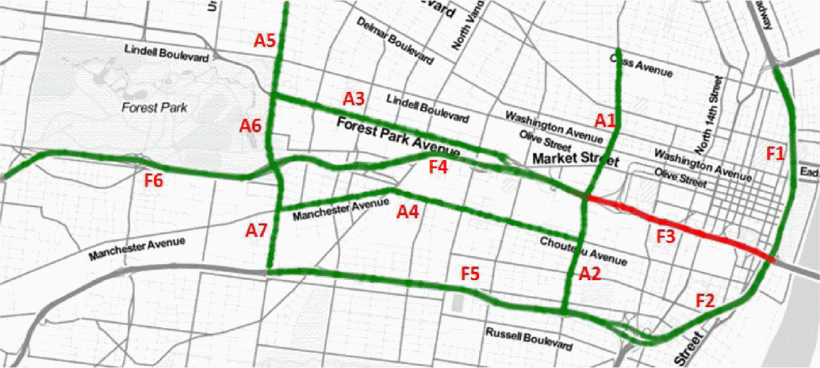


Figure 2: The traffic network (Hou et al., 2017)

The historical travel time of F3(in red) and the other roads in the same area (in green) were used to train the Random Forest Model. Bagging method was used to build the test-sets. The out-of-bag dataset was used to assess the model. In order to evaluate this Random Forest model, the authors perform the same experiment on a Baseline Predictor model and another two models named: Gradient Boosting Machines (GBM) and AdaBoost. They were then compared with the same criteria: mean absolute percentage error (MAPE).

After this, the Random Forest model was applied to the complex real problem in real-time, predicting the network scale road travel time. This time, they compared Random Forest model with GBM model and another two models: long short-term memory (LSTM) and convolutional neural network (CNN). The results showed that Random Forest can outperform its rivals in many aspects.

## 3.3 Advantages of Random Forests technique on Travel Time Prediction

The advantages of using Random Forest method in travel time prediction can be list as below:

1. Accuracy in single road prediction: Random Forest's performance can be highly optimized by parameter selection. For example, in the single road prediction test, after setting the tree depth: size of the terminal node as 5, number of variables as 18, number of the tree about 100, this model reach the top accuracy. The test result indicates that the Random Forest had less mean absolute percentage error (MAPE) than all its opponents: Baseline Predictor, GBM and AdaBoost.

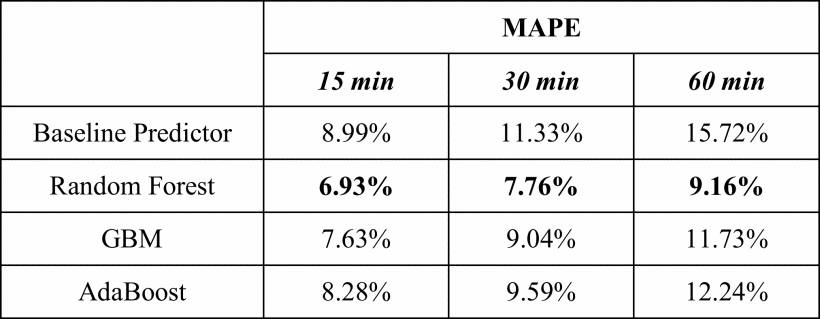


Table 1: TRAVEL TIME PREDICTION RESULTS FOR 15-MIN, 30-MIN, AND (Hou et al., 2017)

1. Accuracy in 13-roads traffic network: When applying to a 13-roads network, Random Forest model still outperform the Baseline and GBM model. Another two deep learning model CNN and LSTM were added for comparison. These two methods cans are parallel training models, they can calculate these 13 roads at the same time. However, they still can win Random Forest.

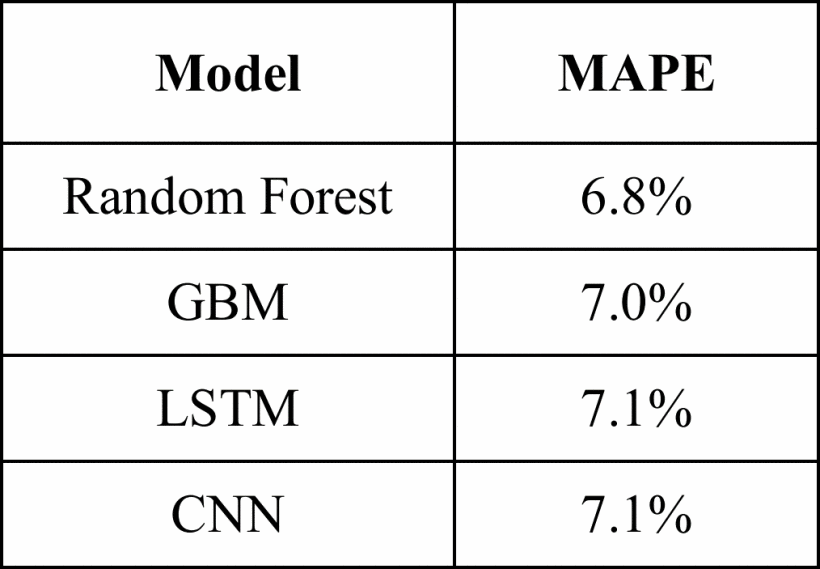


Table 2: COMPARISON WITH DEEP LEARNING (Hou et al., 2017)

1. Computation times: Random Forests are very quick to train and to predict. Table 3 shows that Random Forest model finished training and gave prediction in a relatively short time, although the computational test was run on a not very high-power computer.



Table 3: COMPUTATION TIMES (Hou et al., 2017)

1. Feature Engineering: After the training, Random Forest can reveal that which features are more important. The importance is calculated by Gini index. It is very helpful for a human to understand the relationship between all kinds of constraints and target. Fig 3 illustrates that the historical variables of road F6, A6, F4, F2 and F3 itself are very important in predicting the F3's travel time in the 1-hour future.

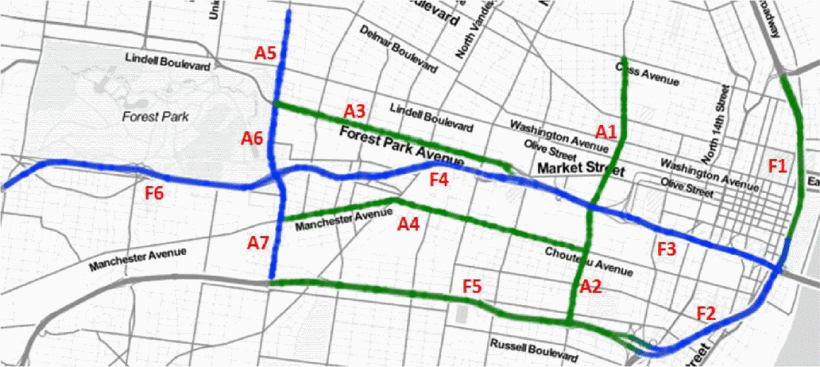


Table 4: The most relevant segments for 1 hour ahead of F3 (Hou et al., 2017)

1. Random Forest can easily use the out-of-bag datasets for validation. Thus, no extra workload is needed to build the test set for Cross-Validation.
2. Random Forest is not sensitive to missing data. In this application, 0.5% of data were missing, they were replaced by mean values. This kind of approach does not affect the result accuracy.
3. The Random Forest model in this application was sensitive to emergency and abnormal condition. It can predict the abnormal congestion of the target road according to the other roads historical travel times.

## 3.4 Disadvantages of Random Forests technique on Travel Time Prediction

The using Random Forest on Road Network travel time estimation had several disadvantages.

1. The big noise in some Classifier may lead to overfitting. Therefore, the trial and error approach were adopted to find out the best depth of the trees or the minimal size of nodes.
2. Second, Random Forest model cannot explain why there were spatial effects. Some travel time of the roads far away from the target road can have more significant influence on the prediction. While the data of some neighbour roads had less impact.
3. Third, how the continuous values were split into different classes may have a significant impact on the structure of Random Forest. This may be the reason why freeway segments obtained less accuracy compared to signalized arterials.

## 3.5 Other available techniques that can be used for the travel time prediction

Other available techniques such as gradient boosting method (GBM) and AdaBoost algorithm (AdaBoost) were studied to improve the prediction.

Not like Random Forest, GBM technique does not average the results of trees in the forests. Instead, it builds the tree models one by one. The latter model fixes the error brought by the preceding model. This building process is under the guidance in order to obtain a better prediction (Zhang & Haghani, 2015) .

The AdaBoost algorithm is a famous algorithm for boosting methods. It took Random Forest as a week learner. In the training process, more weight was set to the training sets which have a higher error rate. In the end, it combines all the learner into the powerful learner a is the final model. The disadvantage of AdaBoost is that the abnormal data may accumulate it weight then finally affect the accuracy (Leshem & Ritov, 2007) .

# 4. Future Development and Conclusion:

## 4.1 The future of the technique

In the future, more and more improved Random Forest model will be proposed, along with many high efficient and free libraries in R and Python languages will be developed. People can then easily adopt these outstanding models to settle their problem.

## 4.2 The future of the technique in relate to ABI

Now, Random Forests already be used to identify some stock’s behaviour, in the future it could be used to uncover the stock trends.

Random Forests can be used to identify and make a good decision to deal with the environment problem, such as ground-level PM2.5 concentration recognition (Liu et al., 2018) .

## 4.3 Conclusion:

In conclusion, Random Forest is a simple, straight-forward but effective technique for capturing knowledge from information. It has many outstanding advantages and has been studied and improved to a great extent. Although it still has some drawback in some areas, this will not stop the data scientists to put more effort into it.

References

Akansha, S. (2015). How to create A perfect decision tree. Retrieved from https://www.edureka.co/blog/decision-trees/

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Genuer, R., Poggi, J., Tuleau-Malot, C., & Villa-Vialaneix, N. (2017). Random forests for big data. *Big Data Research, 9*, 28-46.

Hou, Y., Edara, P., & Chang, Y. (2017). (2017). Road network state estimation using random forest ensemble learning. Paper presented at the *Intelligent Transportation Systems (ITSC), 2017 IEEE 20th International Conference On,* 1-6.

Hu, C., Chen, Y., Hu, L., & Peng, X. (2018). A novel random forests-based class incremental learning method for activity recognition. *Pattern Recognition, 78*, 277-290.

Kim, K., & Giles, C. L. (2016). Financial entity record linkage with random forests. Paper presented at the *Proceedings of the Second International Workshop on Data Science for Macro-Modeling,* 13.

Leshem, G., & Ritov, Y. (2007). Traffic flow prediction using adaboost algorithm with random forests as a weak learner. Paper presented at the *Proceedings of World Academy of Science, Engineering and Technology, , 19* 193-198.

Liu, Y., Cao, G., Zhao, N., Mulligan, K., & Ye, X. (2018). Improve ground-level PM 2.5 concentration mapping using a random forests-based geostatistical approach. *Environmental Pollution, 235*, 272-282.

Michalewicz, Z., Schmidt, M., Michalewicz, M., & Chiriac, C. (2006). Adaptive business intelligence. *Adaptive business intelligence* (pp. 37-46). Berlin, Heidelberg: Springer.

Rokach, L., & Maimon, o. Z. (2014). Introduction to decision trees. *Data mining with decision trees: Theory and applications* (pp. 1-11) World scientific.

Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies, 58*, 308-324.

Zhang, Y., Cao, G., Li, X., & Wang, B. (2018). Cascaded random forest for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,*