Registration Number: 200226767

Word Count: 2006

Abstract

This report aims to compare ensemble classifiers with non-ensemble classifier through the prediction of whether or not a passenger will survive the sinking of the Titanic. The two ensemble methods used to build the prediction models are XGBoost and Random Forest, which are both decision-tree-based algorithms. The result shows that XGBoost (0.92) has the highest AUC, followed closely by Random Forest (0.90), and Decision Tree (0.84) has the lowest performance. Conclusion: Ensemble approaches have higher performance on the prediction of the titanic data than non-ensemble methods.

Introduction

The purpose of this report is to predict whether passengers survived in the sinking of the Titanic and to compare the performance of ensemble methods with that of non-ensemble method. The target column is called “Survived”, and the value 0 means “Not survived”, and 1 means “Survived”. This also indicates that the prediction will be a classification. The result of “XGBoost” and “Random Forest” will be comparing with the result of “Decision Tree” to see if ensemble methods outperform non-ensemble methods. The Titanic dataset is used on the classification which includes two datasets “personal data” and “ticket data”. These two data will be joined by the column “PassengerId”. Kelleher (2015) demonstrate that getting to know the data first is important. I will use the joined data to do the visualization and explore the characteristic of the attributes and how these attributes relate to the target column. After that, I will process the data by splitting or aggregating certain columns, and deal with missing data. Finally, the processed data will be used in the three classifiers and cross validation will be executed. In the end, I will use ROC curve and AUC to evaluate the models built with each method. ROC is often used to evaluate binary classification (Fong et al.,2016) (Saito et al.,2017). On the other hand, ROC is less likely to be impacted by the distribution (Li et al.,2016) which is very suitable for imbalanced data like the one used in this report (those not survived are almost twice the time as those survived).

Data mining theory

In this section, I will introduce the selected methods and the reason for the choices. After that, I will in describe how the methods work separately. Finally, the choice of evaluation will be explained.

Firstly, the prediction target is to put the passengers into two class “survived” and “not survived”, so I will be choosing from classification methods. The base classifier I choose is Decision Tree, and the ensemble methods used to compare with the base classifier are XGBoost and Random Forest. The concept of ensemble methods is to combine many weak learners to get better results. The study of Dietterich (2000) compared the single decision tree with ensemble trees, and the ensemble trees did have better performance. There are many types of ensemble trees such as boosting, bagging and randomization approaches. In this report I use XGBoost and Random Forest approaches because Banfield et al. (2007) suggested that boosting and random forest approach have higher performance than standard bagging. XGBoost and Random Forest are both widely used in classification. Davagdorj et al. (2020) uses XGBoost to do disease prediction and XGBoost showed high performance on the classification. XGBoost algorithm is applied in the process of automotive manufacturing and performed well on the prediction (Chen et al., 2019). XGBoost also showed well performance in classifying patients with epilepsy and orthopedic auxiliary diagnosis (Torlay et al., 2017) (Li & Zhang, 2020). Random Forest showed excellent performance on predicting the susceptibility of landslide (Chen et al., 2018). Random Forest is also applied to support the diagnosis of diseases and shows good performances (Gray et al., 2013) (Dimitriadis & Liparas, 2018). Therefore, I selected these two methods to see whether they outperform decision tree and to what extent.

XGBoost is short for eXtreme Gradient Boosting and it is one of the boosting methods. The weak classifiers in boosting are non-parallel. Every time a new tree is created, the residual of the previous tree is learned, and the weights of the misclassified data is added to train and improve the new tree. In other words, the models are improved through the process of learning the errors made by the previous tree.

Random Forest consists of many parallel trees, and there is no relationship between the trees. Random sampling is applied to the data when training the

decision trees. Random subsets of the features are considered for the node splitting. Finally, the majority voting makes the final prediction. For example, if most trees vote for class A, and few votes for class B, then class A will be the final prediction.

After the models are built, I will use ROC and AUC to do the evaluation. The models in this report are binary classifiers. ROC has been applied to evaluate the performance of classifiers (Fong et al., 2016). Besides, it is especially popular in binary classification (Saito & Rehmsmeier, 2017). On the other hand, the data used in this report is fairly imbalanced for those who did not survived are 63% of the total passengers. In this case, the accuracy of the model might not be as suitable as ROC because of the imbalanced distribution. ROC has also been used to evaluate the classification of imbalanced data (Li et al., 2016), and I believe it would bring value to the evaluation of the models in this report as well.

Data exploration and preparation

Table 1 contains the definition and some notes for the attributes in the Titanic dataset. I will introduce which of the attributes will be used to train the model and the explain the reason.

Variable	**Definition**	**Note**
PassengerId	Weather survived	the identifier
Name	Name of the passenger	First name + Title + Last name
Sex	Gender	Male/ Female
Age	Age	
SibSp	Number of siblings and spouse	
Parch	Number of parents and children	
Salary	Salary in dollars	
Job	Job title	
Survived	Weather survived	0 = Not survived 1 = Survived
Ticket	Ticket Number	

Fare	The passenger fare	
Cabin	Cabin number	
Embarked	Port of embarkation	C = Cherbourg Q = Queenstown S = Southampton

Table 1

Kelleher (2015) suggested to use different ways to get to know categorical data and numerical data. Use bar plots to see the frequencies of categorical data and use mean and standard deviation to understand numerical data. Therefore, I follow Kelleher's suggestion to get to know the data. Kelleher demonstrated that identifiers are not suitable for model training, so PassengerId is not included. I extract the "Title" from the "Name" column. As you can see in Table 2, there are some titles which have the same meaning such as "Miss" and "Mlle". They will be combined together as "Miss", and "the Countess", "Lady" and "Sir" are combined together as "Royalty".

Row ID	S Survived	S Title	I PassengerId...
Row0	0	Capt	1
Row1	0	Col	3
Row11	1	Col	1
Row2	0	Don	1
Row12	1	Dona	1
Row3	0	Dr	5
Row13	1	Dr	3
Row4	0	Jonkheer	1
Row14	1	Lady	1
Row5	0	Major	1
Row6	0	Master	37
Row15	1	Master	22
Row7	0	Miss	48
Row16	1	Miss	191
Row17	1	Mlle	2
Row18	1	Mme	1
Row8	0	Mr	626
Row19	1	Mr	74
Row9	0	Mrs	21
Row20	1	Mrs	153
Row21	1	Ms	1
Row10	0	Rev	8
Row22	1	Sir	1
Row23	1	the Countess	1

Table 2

For “Sex” attribute, we can see from Figure 1,2 that the survived for female is much higher than that of male. If we calculate the number, we’ll see that the survival rate of male is 0.13, and the survival rate of female is 0.83, which suggests that “Sex” could be an influential attribute in this prediction.

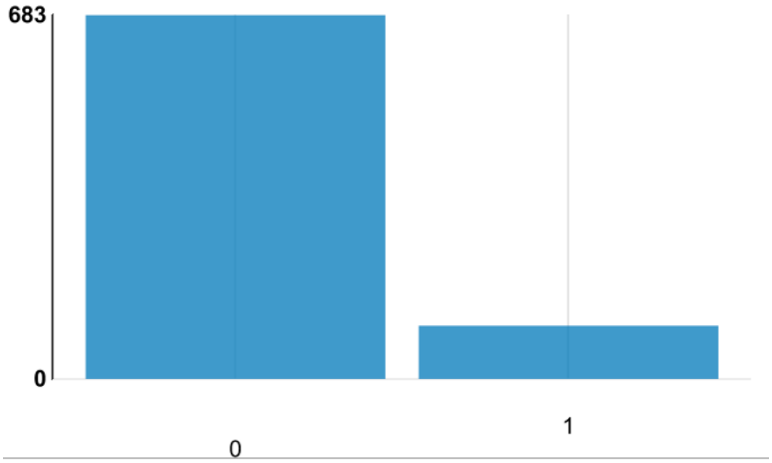

Figure 1 Male survived

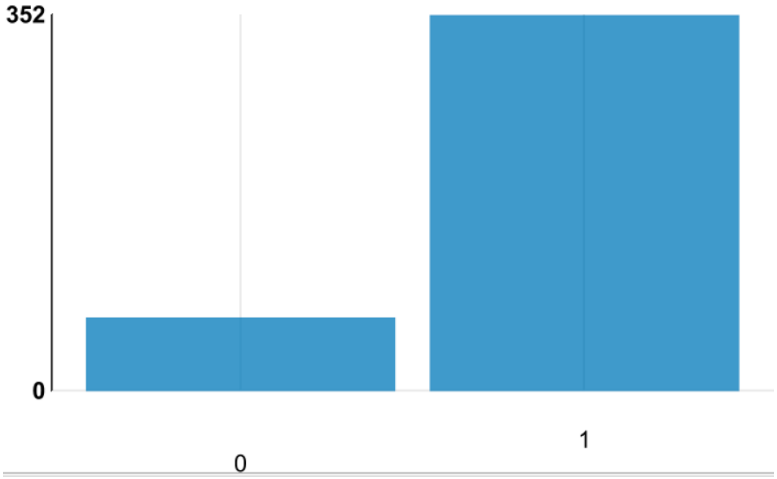

Figure 2 Female survived

Figure 3 shows that most of the passenger's age fall on the range between 21 and 39, and the oldest passenger was 76 years old. I will use mean imputation to replace the 242 missing values of "Age".

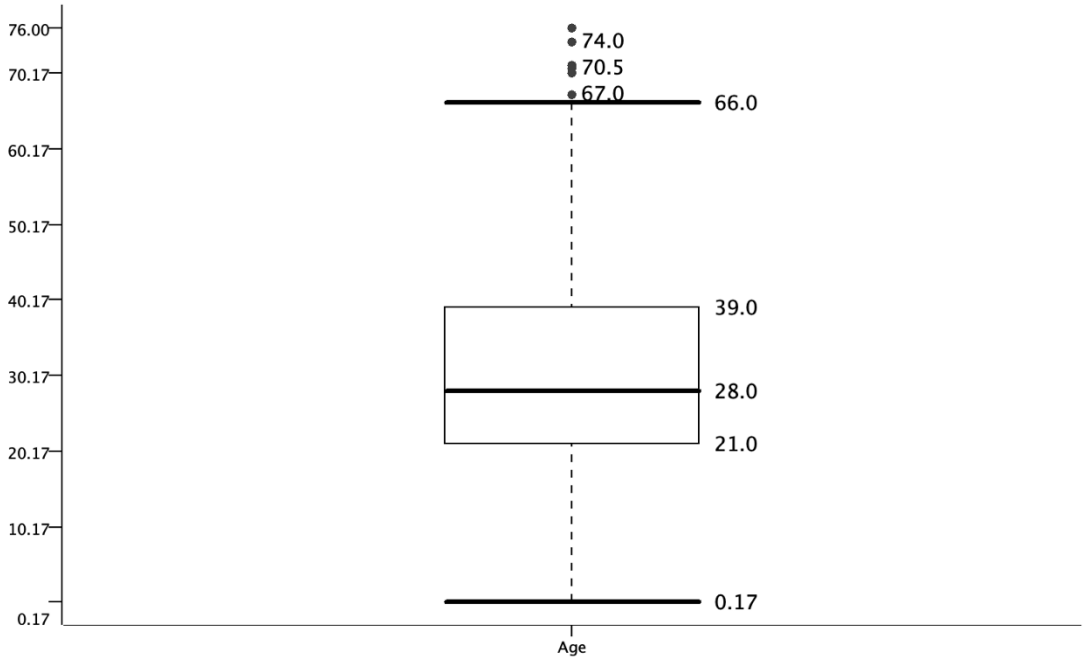

Figure 3

I will combine the number of "SibSP" and "Parch" and create a new column "Family Number". Figure 4 shows the distribution of family number, and most passengers have 0 or 1 family member on Titanic.

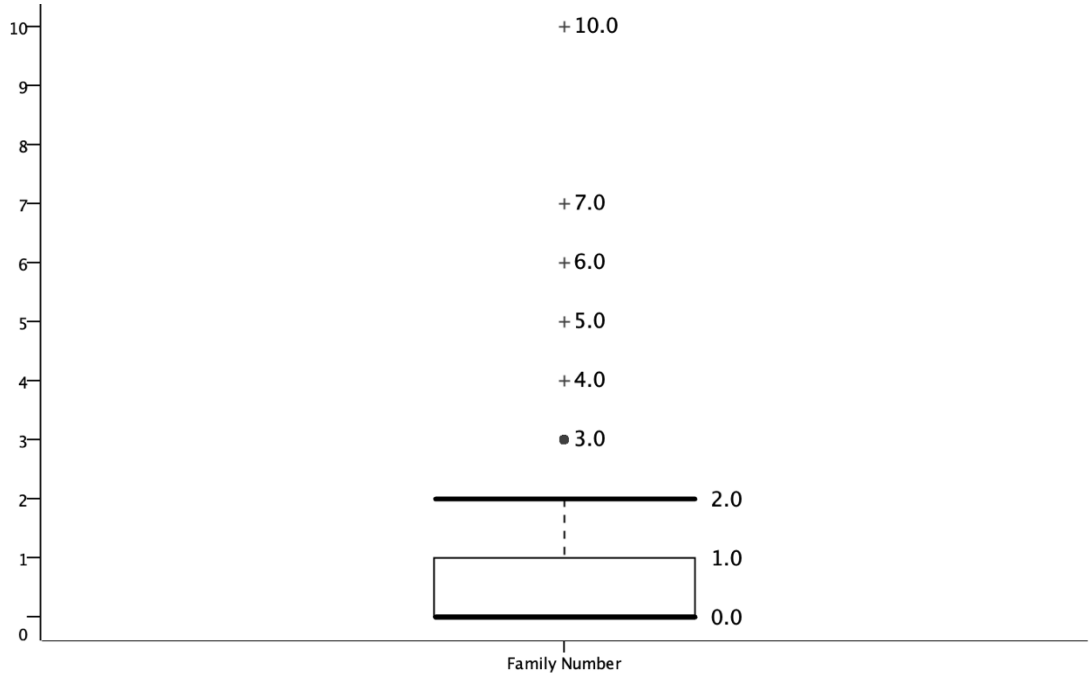

Figure 4

Table 5 shows the distribution of Salary.

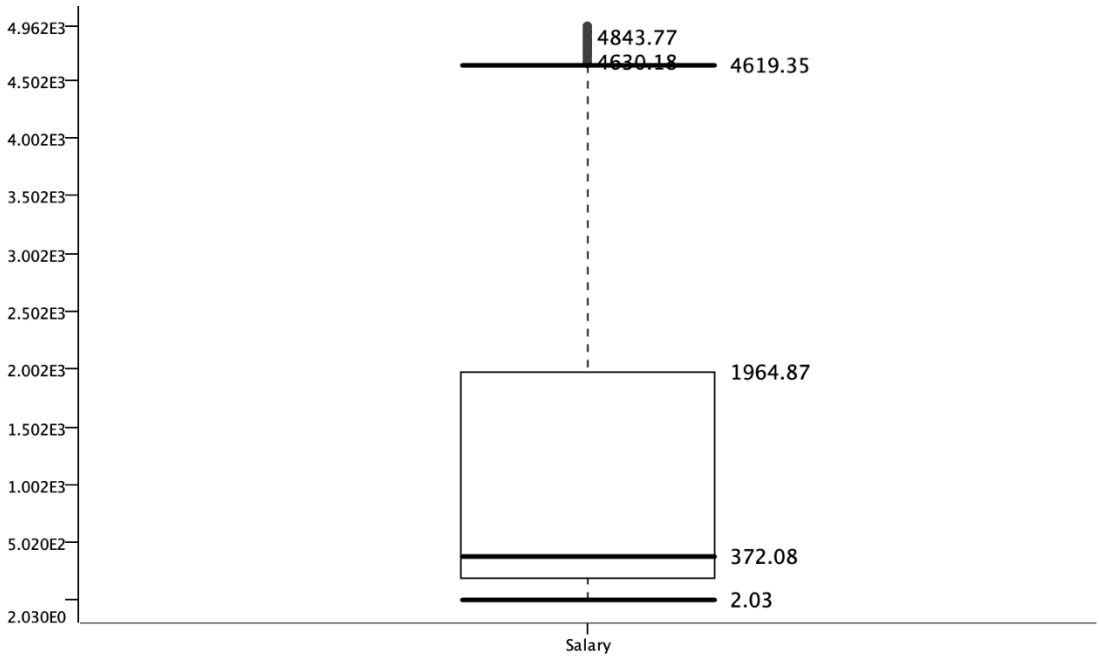

Figure 5

Table 3 shows the survival rate of each job type. What's very interesting is that those medical field jobs have survival rate higher than 0.5. Therefore, I decided to put the jobs in two categories "medical" and "not medical" which indeed slightly improved the model.

Job	**Survival rate**
Apothecary	**0.65**
Pharmacist	**0.54**
Doctor	**0.5**
Surgeon	**0.52**
Dentist	**0.6**
Wigmaker	0.29
Blacksmith	0.29
Barber	0.15
Wheelwright	0.26
Shopkeeper	0.28
Shoemaker	0.29
Hatter	0.5
Breechesmaker	0.48
Cutler	0.4

Cabinetmaker	0.57
Farrier	0.54
Bookbinder	0.42
Carpenter-joiner	0.44
Coach maker	0.48
Goldsmith	0.64
Music teacher	0.59
Silversmith	0.23
Weaver	0.31

Table 3

Figure 6 shows the distribution of “Fare”, ranging from 512.33 to 7.9.

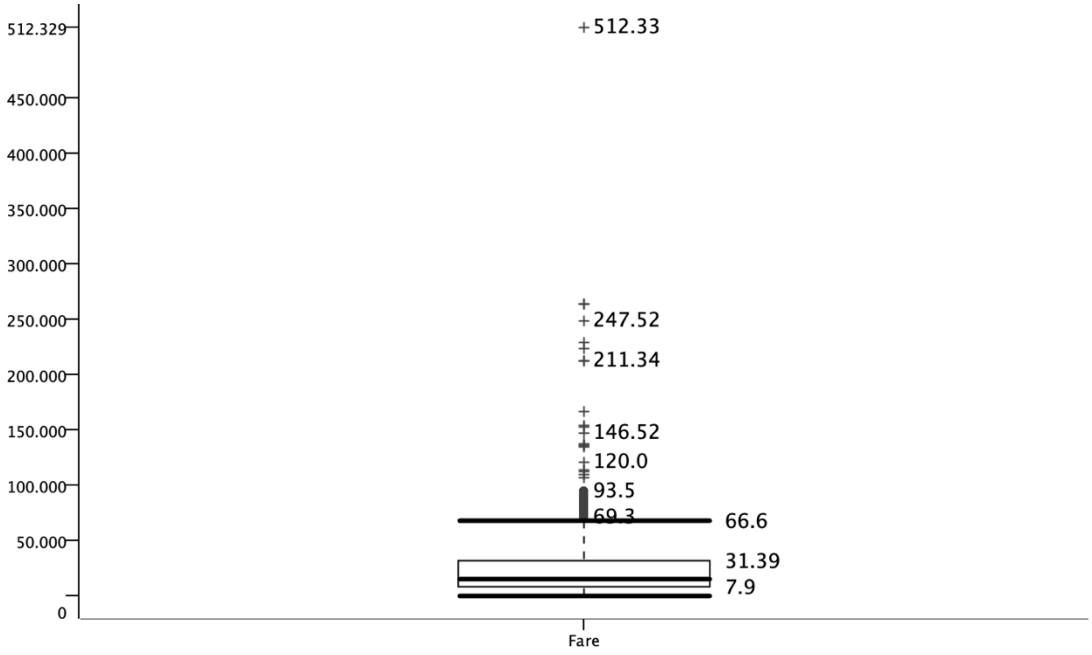

Figure 6

There are 934 missing values in the category “Cabin”, which means 77% of the data is missing. According to Kelleher (2015), when 60% of the data is missing, the attribute should not be used to build the models. However, I found that 60% of the passenger with cabin survived while only 30% of those without cabin survived. Therefore, I put them into two new categories “has cabin” and “no cabin”.

Row ID	Survived	Cabin	PassengerI...
Row0	0	Has cabin	103
Row2	1	Has cabin	167
Row1	0	no cabin	649
Row3	1	no cabin	285

Table 4

Figure7 shows that most people embarked at Southampton, and Table 5 shows that people embarked at Cherbourg are more likely to survive. There are 2 missing values for this category, and they will be replaced by the mode “S” which is suggested by Kelleher (2015).

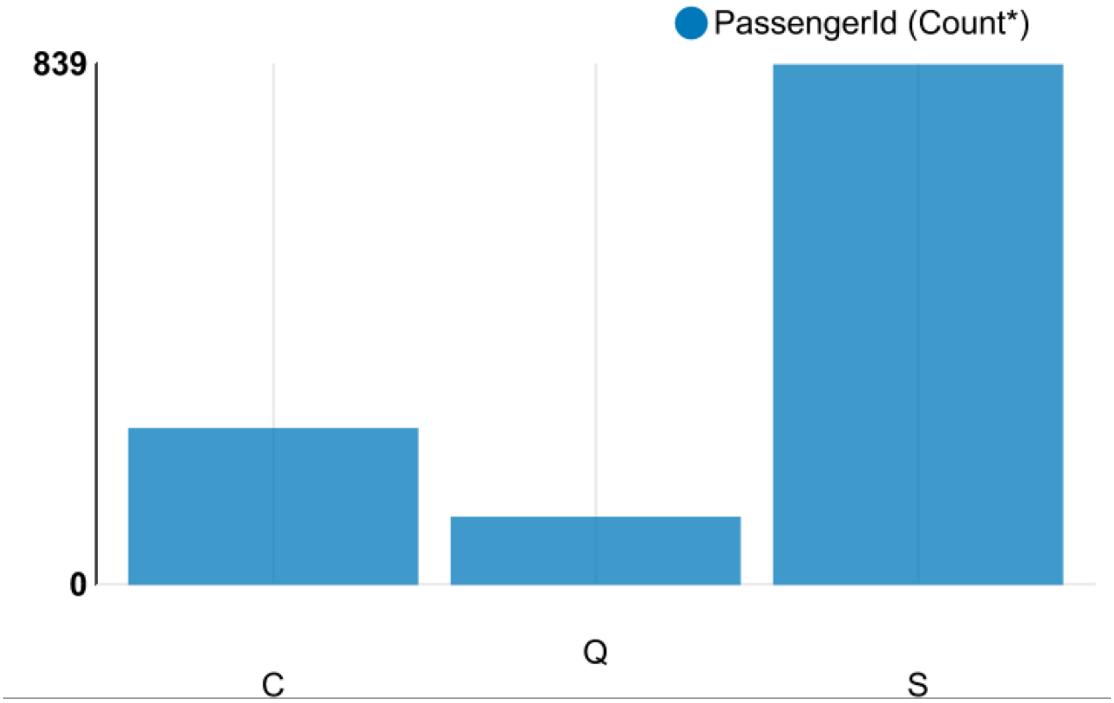

Figure 7

Embarked	Survival rate
C	0.49
Q	0.45
S	0.32

Table 5

Experimental setup

Decision tree

In the decision tree learner node configuration, I choose GINI index as quality measure. GINI index is widely used on splitting nodes for decision trees and has brought value to researches (Mathan et al., 2018) (Chandra & Paul, 2009) (Reddy, 2017). Furthermore, overfitting is a common issue found in decision tree. It lowers the efficiency and reduces the accuracy of models. Therefore, I use MDL-based pruning method to avoid overfitting. MDL-based pruning has been demonstrated to increase efficiency and accuracy on modeling (Pham & Afify, 2006).

Random Forest

For random forest learner node, I use GINI index on the split criterion as well. The number of models is set to be 50.

XGBoost

For XGBoost ensemble tree learner node, I use 50 boosting rounds so the model can learn enough from errors. I also tried 60, 80 and 100 boosting trees, but it did not get better results.

Partitioning

Eighty percent of the data will be used to train the model, and 20% of the data will be used on testing the model.

Cross Validation

First of all, I have used cross validation on all of the three approaches. Cross validation allows us to use different combination of data to validate the model and to avoid relying on only one subset which might cause bias. Cross validation is widely used on model evaluation and selection (Geisser, 1975). It has brought high efficiency to some areas such as application of engineering design (Kang et al., 2019). In Knime, it can be done through the "X-Partitioner" and "X-Aggregator" nodes. For the configuration, I use stratified

sampling on the column for "X-partitioner" node to maintain the class distribution of "Survived" column. For the "X-Aggregator" column, the target is configured as "survived" and the prediction is configured as "Prediction (Survived)".

Results and Discussion

In this part, I will introduce what I have done to improve the models and the ROC and AUC of each method to compare their performance. After that, I will discuss which methods outperform and why they have better performance. Model improvement in this report is conducted through boosting, feature selection and pruning. Boosting is used on XGBoost method, pruning is used on decision tree and feature selection is used on all of them. ROC and AUC tell us how much the model can classify the two classes. Figure 8 is the ROC curve of Decision Tree. The AUC is 0.84 which is higher than the model without pruning (0.80).

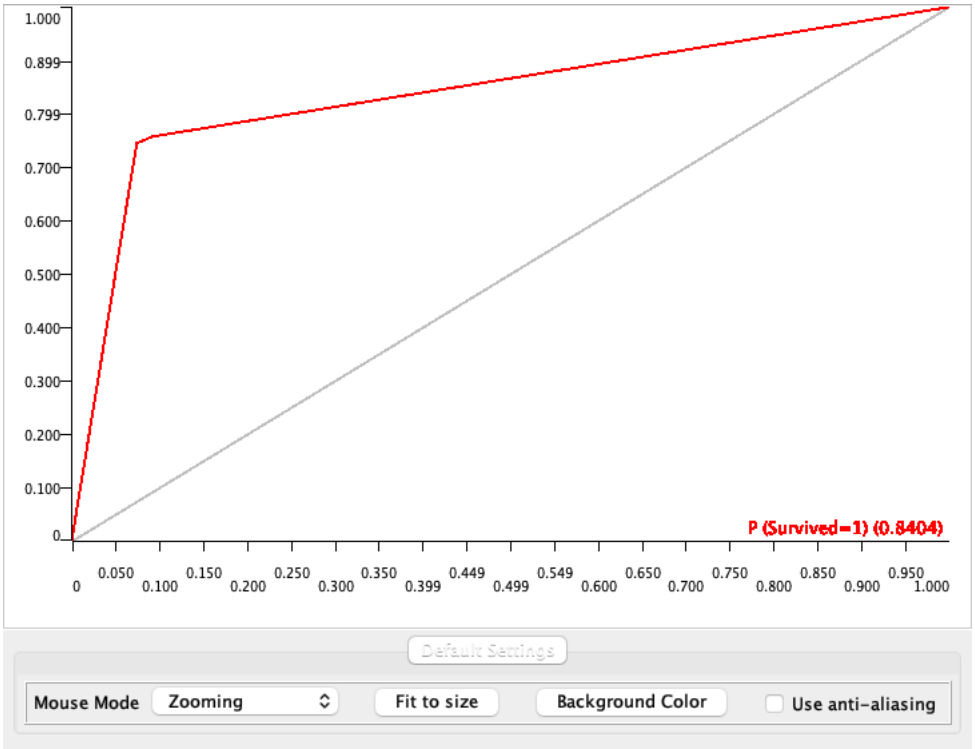

Figure 8 - ROC of Decision Tree

Figure 9 is the decision tree view which is pruned. Therefore, only three of the attributes were included in this view. The first split is “Sex” which indicates that “Sex” attribute is the most important attribute in the prediction of Titanic sinking. “Family number” and “Salary” for female are the second important attributes.

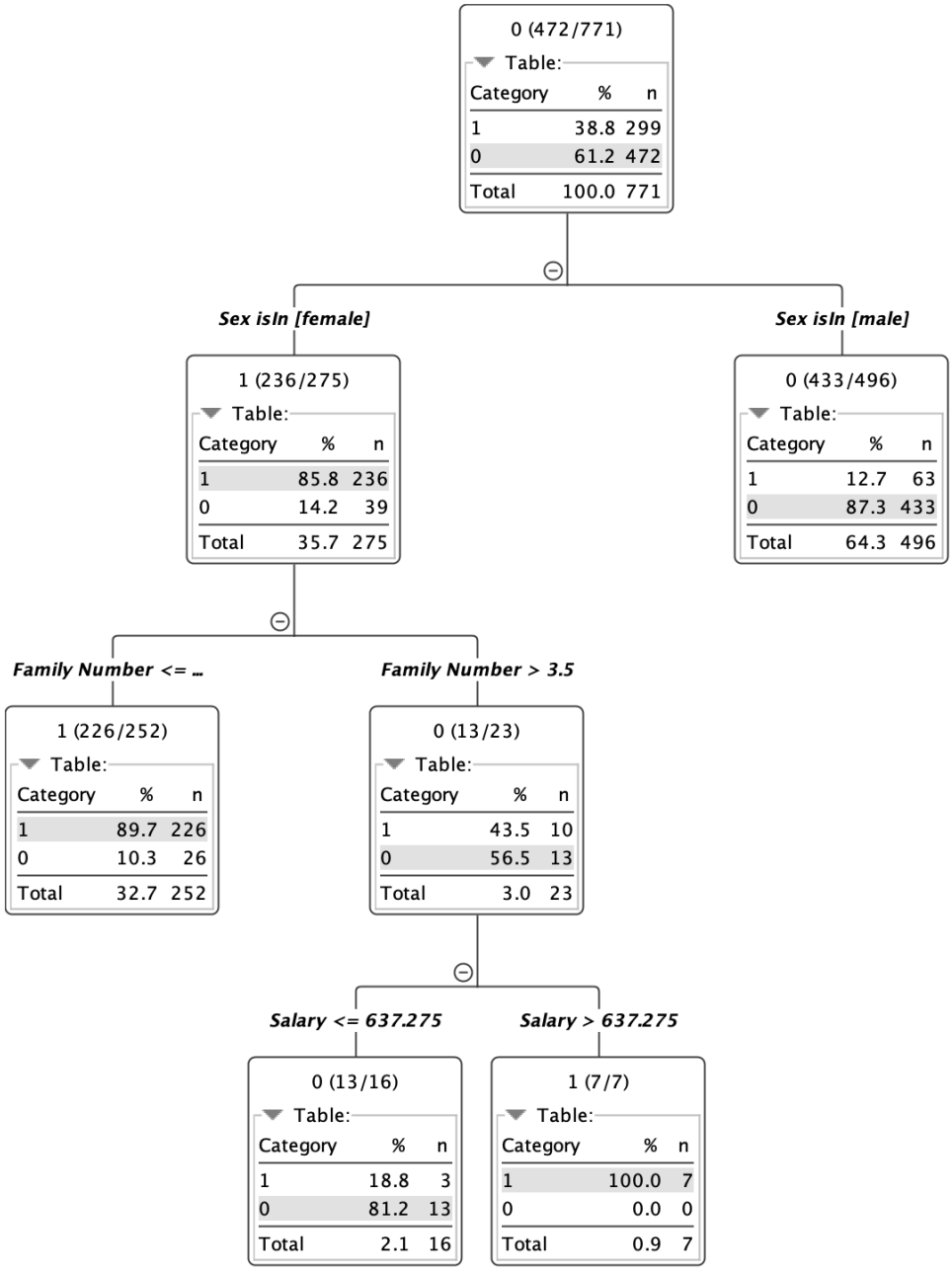

Figure 9 - Decision Tree View

Figure 10 is the ROC of Random Forest. It has the highest AUC 0.90.

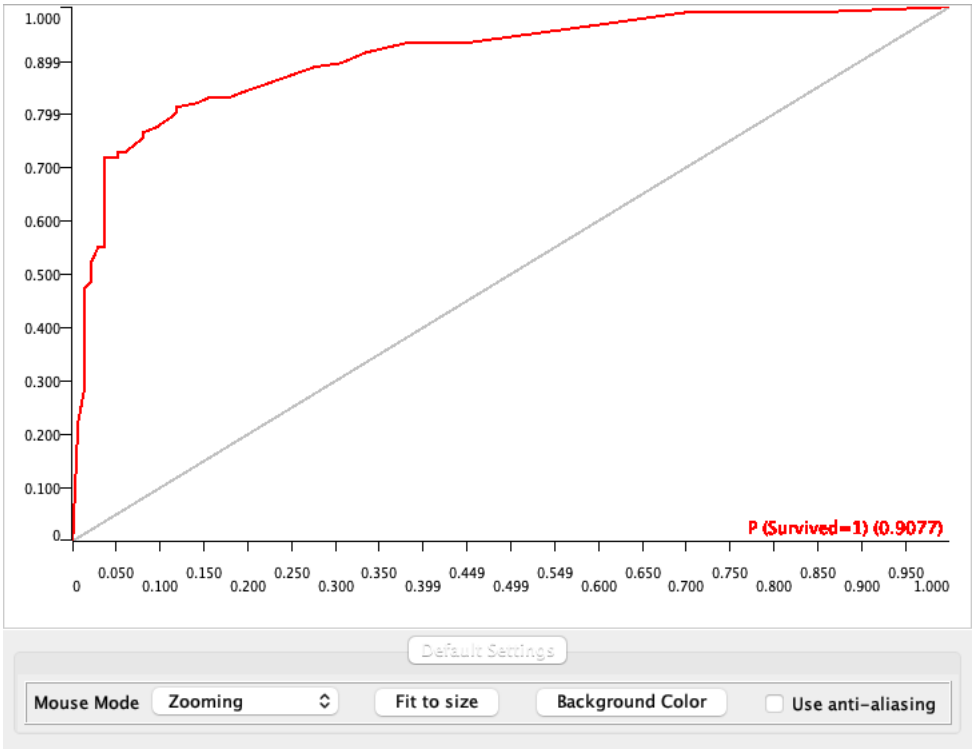

Figure 10 - ROC of Random Forest

Figure 11 is the ROC curve of XGBoost, and the AUC is 0.92.

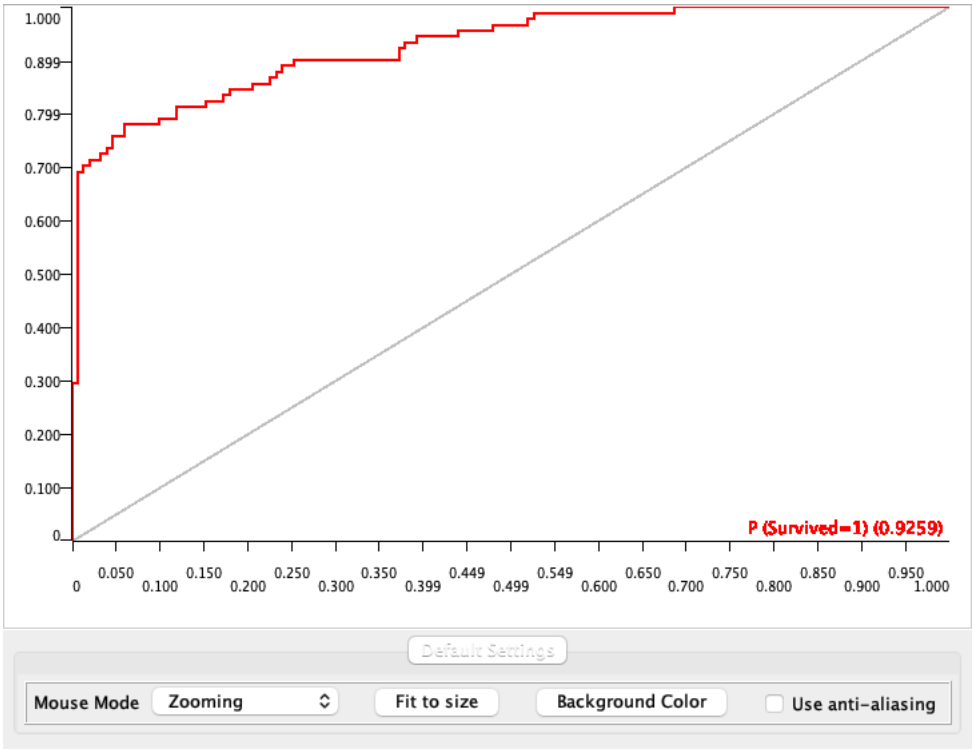

Figure 11 - ROC of XGBoost

The AUC of XGBoost (0.92) and Random Forest (0.90) are both higher than that of Decision Tree (0.84). We can tell from the result that the ensemble trees do perform better than single tree. The evaluation shows that Random Forest performs better than Decision tree model. Random Forest has high accuracy, no overfitting problem and is robust to the noise. As the trees grow deeper in Random Forest, the generalization error meets a limit and improves the model and thus make it better than a single tree (Sekhar & Mohanty, 2016). The votes of the trees in Random Forest makes the model less noisy and less sensitive to outliers which then improves the robustness (Shaikhina et al., 2019). The performance of XGBoost is also better than Decision tree. XGBoost reduces the time it takes to grow trees, which means it is fast (Chen & Guestrin, 2016). The error declines when new iteration learn from the residuals and misclassified samples from the previous iteration. Hence, it gets more reliable results (Basak et al., 2019).

Conclusion and reflection

In the Titanic sinking prediction, Random Forest and XGBoost both have better performance than Decision Tree. Random Forest lowers the impact of data bias (Shaikhina et al., 2019), and XGBoost learns from the previous trees to improve the models (Basak et al., 2019) which is why they can do better than Decision Tree. And of all the three approaches, XGBoost shows the best performance in this prediction.

One improvement can be done in this report is to reduce the influence that imbalance data causes. Of all the three models, the accuracy of predicting “survived” is about 0.7, and nearly 0.9 for “not survived”. Many researches have suggestions over this issue, such as oversampling (Liang, Li & Hu, 2018) and undersampling (Yin et al., 2014), data reduction and stacking (Czarnowski & Jedrzejowicz, 2019).

References

Banfield, R.E, Hall, L.O, Bowyer, K.W, & Kegelmeyer, W.P. (2007). A Comparison of Decision Tree Ensemble Creation Techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1), 173-180.

Basak, Suryoday, Kar, Saibal, Saha, Snehanshu, Khaidem, Luckyson, & Dey, Sudeepa Roy. (2019). Predicting the direction of stock market prices using tree-based classifiers. *The North American Journal of Economics and Finance*, 47, 552–567. https://doi.org/10.1016/j.najef.2018.06.013

Chandra, B, & Paul Varghese, P. (2009). Fuzzifying Gini Index based decision trees. *Expert Systems with Applications*, 36(4), 8549–8559. https://doi.org/10.1016/j.eswa.2008.10.053

Chen, Wei, Zhang, Shuai, Li, Renwei, & Shahabi, Himan. (2018). Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling. *The Science of the Total Environment*, 644, 1006-1018.

Chen, Kai, Chen, Huabin, Liu, Liang, Chen, Shanben. (2019). Prediction of weld bead geometry of MAG welding based on XGBoost algorithm. *International Journal of Advanced Manufacturing Technology*, 101(9), 2283-2295.

Chen, Tianqi, & Guestrin, Carlos. (2016). XGBoost: *A Scalable Tree Boosting System*. https://doi.org/10.1145/2939672.2939785

Czarnowski, Ireneusz, & Jedrzejowicz, Piotr. (2019). Data reduction and stacking for imbalanced data classification. *Journal of Intelligent & Fuzzy Systems*, 37(6), 7239–7249. https://doi.org/10.3233/JIFS-179335

Davagdorj, K., Pham, V. H., Theera-Umpon, N., & Ryu, K. H. (2020). Xgboost-based framework for smoking-induced noncommunicable disease prediction. *International Journal of Environmental Research and Public Health*, 17(18), 1–22. https://doi.org/10.3390/ijerph17186513

Dietterich, T.G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Mach. Learn*. 2000, 40, 139–157.

Dimitriadis, Stavros, & Liparas, Dimitris. (2018). How random is the random forest? Random forest algorithm on the service of structural imaging biomarkers for Alzheimer's disease: From Alzheimer's disease neuroimaging initiative (ADNI) database. *Neural Regeneration Research*, 13(6), 962-970.

Fong, Youyi, Yin, Shuxin, & Huang, Ying. (2016). Combining biomarkers linearly and nonlinearly for classification using the area under the ROC curve. *Statistics in Medicine*, 35(21), 3792-3809.

Geisser, Seymour. (1975). The Predictive Sample Reuse Method with Applications. *Journal of the American Statistical Association*, 70(350), 320-328.

Gray, Katherine R, Aljabar, Paul, Heckemann, Rolf A, Hammers, Alexander, & Rueckert, Daniel. (2013). Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *NeuroIma*ge *(Orlando, Fla.)*, 65, 167-175.

Kelleher, J. D. (2015). Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. *The MIT Press*.

Kang, Kyeonghwan, Qin, Caiyan, Lee, Bongjae, & Lee, Ikjin. (2019). Modified screening-based Kriging method with cross validation and application to engineering design. *Applied Mathematical Modelling*, 70, 626-642.

Li, Yang, Qin, Yichen, Wang, Limin, Chen, Jiaxu, & Ma, Shuangge. (2016). Grouped Variable Selection Using Area under the ROC with Imbalanced Data. *Communications in Statistics. Simulation and Computation*, 45(4), 1268–1280. https://doi.org/10.1080/03610918.2013.818691

Li, Shenglong, & Zhang, Xiaojing. (2020). Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm. *Neural Computing & Applications*, 32(7), 1971-1979.

Liang, Peifeng, Li, Weite, & Hu, Jinglu. (2018). Oversampling the minority class in a multi-linear feature space for imbalanced data classification. *IEEJ Transactions on Electrical and Electronic Engineering*, 13(10), 1483–1491. https://doi.org/10.1002/tee.22715

Mathan, K, Kumar, Priyan Malarvizhi, Panchatcharam, Parthasarathy, Manogaran, Gunasekaran, & Varadharajan, R. (2018). A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease. *Design Automation for Embedded Systems*, 22(3), 225–242. https://doi.org/10.1007/s10617-018-9205-4

Pham, D T, & Afify, A A. (2006). Three New MDL-Based Pruning Techniques for Robust Rule Induction. Proceedings of the Institution of Mechanical Engineers. Part C, *Journal of Mechanical Engineering Science*, 220(4), 553–564. https://doi.org/10.1243/09544062C18404

Reddy, S.V.G. (2017). ENHANCING THE SPEED, ACCURACY OF DEEP LEARNING USING GINI INDEX BASED FUZZY DECISION TREES. *International Journal of Advanced Research in Computer Science*, 8(9), 411–417. https://doi.org/10.26483/ijarcs.v8i9.5053

Saito, Takaya, & Rehmsmeier, Marc. (2017). Precrec: Fast and accurate precision-recall and ROC curve calculations in R. *Bioinformatics*, 33(1), 145-147.

Shaikhina, Torgyn, Lowe, Dave, Daga, Sunil, Briggs, David, Higgins, Robert, & Khovanova, Natasha. (2019). Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomedical Signal Processing and Control*, 52, 456–462. https://doi.org/10.1016/j.bspc.2017.01.012

Sekhar, Pudi, & Mohanty, Sanjeeb. (2016). Classification and assessment of power system static security using decision tree and random forest classifiers. *International Journal of Numerical Modelling*, 29(3), 465–474. https://doi.org/10.1002/jnm.2096

Torlay, L, Perrone-Bertolotti, M, Thomas, E, & Baciu, M. (2017). Machine learning–XGBoost analysis of language networks to classify patients with epilepsy. *Brain Informatics*, 4(3), 159-169.

Yin, Qing-Yan, Zhang, Jiang-She, Zhang, Chun-Xia, & Ji, Nan-Nan. (2014). A Novel Selective Ensemble Algorithm for Imbalanced Data Classification Based on Exploratory Undersampling. *Mathematical Problems in Engineering*, 2014, 1–14. https://doi.org/10.1155/2014/358942