CE888 DATA SCIENCE AND DECISION MAKING Assignment-2

Project 3: Knowledge distillation from random forests (Supervised learning)

Jayakrishnadeva Suresh Balasundaram - 1802552

Abstract— **Ensemble methods are collection of algorithms or procedures that learns to build a set of classifiers that classifies unknown observations by polling their predictions (weighted). The first and authentic ensemble method is Bayesian averaging, but recent algorithms tend to integrate with other techniques that can improve the performance of the model. The techniques are output coding, bagging and boosting for enhancing the model by correcting the error. The proposed concept in this paper examines these methods and explains why group of classifier can often perform better than any single classifier[4]. RandomForest naturally incorporates ensemble of decision trees and is non - parametric, i.e. the data points are the parameters of the model. It is designed to perform feature selection in the learning process and is highly predictive.**

I. INTRODUCTION

Finance is one of the major sector where machine learning is leveraged to predict outcomes that are beneficial to institutions in multiple ways. Three dataset are chosen in order to apply the techniques specified and gain exposure in the finance sector. The first dataset is explained below,

One of the major obstacle that any existing banking venture faces is to assess the credibility limit of an individual. The traditional way to overcome this is to set standard income level limit and history of transaction. However, many banking institution are failing to meet the right criteria to set right standard in order to utilise a resource or client. Due to this, many data mining tasks require data classification in classes. Refer to table-1 for detailed information. Loan applications, for example, can be classified as ”approve” or ”disapprove.” A class provides a function to map a data item (instance) into one of several pre-defined classes[5].

The second dataset is about credit card issuing and it is explained as follows, Credit card issuers in Taiwan have faced a crisis in cash and credit card debt recently and in the 3rd quarter of the year 2006, delinquency is expected to peak [22] refer table-2 for more information. In order to increase market share, banks issuing cards in Taiwan gave to applicants who are not fit to receive over-issued cash and credit cards. Meanwhile, most of the cardholders, regardless of their ability to repay, exceeded the use of credit cards for consumption. The crisis has influenced confidence in consumer finance and is considered as one of the major challenge for banks. Crisis management will be downstream and risk prediction will be upstream in a developed financial system. One of the main purpose of risk prediction is to use the information collected from the finance records, such as statements, transactions made by the customer and records maintained for repayment, to predict the performance of business or credit risk for individual customers and to reduce damage and uncertainty.

The third dataset is about the marketing strategies followed in the financial sector and explained as follows, Marketing sales campaigns are a typical business enhancement strategy. Companies use direct marketing to target customer segments by contacting them to achieve a specific objective. The centralisation of remote customer interactions in a contact centre facilitates campaign operational management. Such centres allow customers to communicate via various channels, and one of the most widely used is the telephone (fixed line or mobile). Business Intelligence is an umbrella term that includes architecture & tools, applications, methodologies, and DB in order to use data to support decision making process with business managers. Data Mining is a Business Intelligence technology that includes useful knowledge such as patterns from complex and vast data by modelling the data[20]. Due to the remoteness characteristic [9], marketing operationalized by a contact centre is called telemarketing.

II. BACKGROUND

We briefly review the work done in the past for the chosen data set, decision trees and knowledge distillation,

Decision - tree[8] is usually constructed through partitioning recursively. For selecting the root of the entire tree, a single attribute is nominated using certain criteria. The criteria might include but not limited to gini index, gain ratio, information. The data are then divided with respect to their test and each and every child repeats the process recursively. A step for pruning is performed once the full tree is built. This reduces the overall size of the tree that is produced. In our experiments, we compare our results with the algorithm for decision-making tree C4.5 (Quinlan 1993), which is a state of the art algorithm.[7] In the century of information outburst, each and every day large volumes of data are produced and stored by individual companies. It is a major challenge for companies to identify useful data and knowledge from the data and transform the data into information and actionable results. Data mining is a process to automatically or semi - automatically explore and analyse large amounts of data to discover relevant patterns and rules[22]. At present, data mining is an essential tool for decision support and plays a key role in market segmentation, fraud detection, customer services, credit and behaviour scoring and bench-marking (Paolo, 2001, Thomas, 2000).

III. Methodology

The analysis on three different dataset will provide us results on whether ensemble methods are performing better than machine learning methods. The dataset chosen are related to finance and banking. All the dataset are chosen with a view to provide a classification output from the prediction. The dataset are described in a tabular format,

TABLE I
ADULT DATA DESCRIPTION

Attribute names	Type	Description
Age	continuous	Age of the individual living in the region
Work class	Categorical	Class of the employee according to the work nature
Education	categorical	Attained education status at the time of collecting data
Education-num	continuous	Number of years of Education at the time of collecting data
Marital-status	categorical	Relationship status explained - if married
Occupation	categorical	Current Designation at the organization where the individual works
Relationship	categorical	Current Relationship status at the time of collecting the data
Race	categorical	Ethnic origin of the individual involved
Sex	categorical	Gender of the observation
Capital-gain	continuous	Gain obtained with capital through investment
Capital-loss	continuous	Loss obtained with capital through investment
Hours-per-week	continuous	Number of hours working at the time of collecting the data
Native-country	categorical	Country they belong to or the nationality of the person

The table-1 dataset holds information about salary along with census data. This data is Extracted from the 1994 Census database by Barry Becker. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) and (AGI>100) and (AFNLWGT>1) and (HRSWK>0)). The original motive and the Prediction task is to determine whether a person makes over 50K a year but here we are using this dataset to calculate and compare the performance between ensemble models and normal models as the problem of dataset is considered as a classification problem.

The research for dataset described in table-2 was originally aimed at the case of clients default payment in Taiwan and compares the probability of predictive accuracy of defaulters among many data mining methodology. From the point of view of risk management, the results of predictive accuracy on the calculated probability of defaulters will be more important than the binary result of classification - credible or not credible clients. Since the real probability of default is unknown, the previous study presented the new sorting smoothing method to estimate the real likelihood of default. With the real probability of default as a response variable and the predictive probability of default as an independent variable, the linear regression showed that the predictive model produced by the artificial neural network is having the highest coefficient of determination. Therefore, it was confirmed that among all the data mining techniques, Only artificial neural network can estimate the real probability of defaulters accurately.

TABLE II
DEFAULT OF CREDIT CARD CLIENTS DATASET

Attribute names	Type	Description
X1	Quantitative	Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
X2	Categorical	Gender (1 = male; 2 = female).
X3	Categorical	Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
X4	Categorical	Marital status (1 = married; 2 = single; 3 = others).
X5	Categorical	Age (year).
X6 - X11	Quantitative	History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; 8 = payment delay for eight months; 9 = payment delay for nine months and above.[1]
X12-X17	Quantitative	Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; X17 = amount of bill statement in April, 2005[1].
18-X23	Quantitative	Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005[1].

Few analysis has been performed in the dataset of table-3, where the variable pdays holds an value '999' - This means client was not previously contacted and the variable duration highly affects the target value. For instance if duration=0 then y will be 'no' but, the duration will not be known before a call. Adding to the statement, the y will obviously be known after the call. Thus, this input should only be included only for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

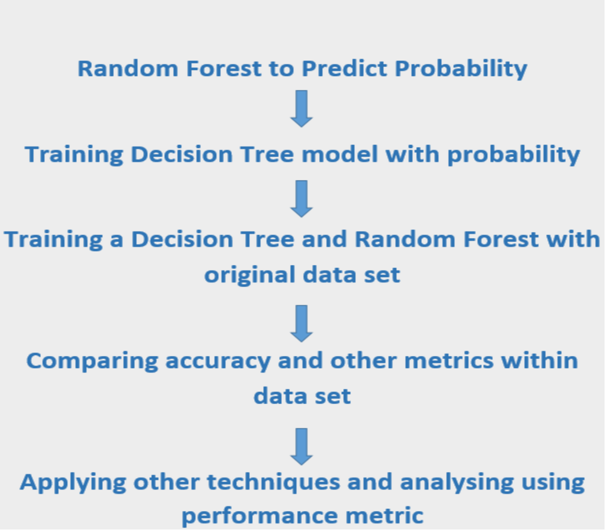
Feature extracting is one of the ways through which we can improve the predicting capability of the model. PCA is one of the feature extracting method through which principle components are created by capturing all the important information from the given data in the first few components. This method can be considered in order to get a best model out of ensemble models.

TABLE III
BANK MARKETING DATA SET

Attribute names	Type	Description
age	Quantitative	Age of the individual living in the region
job	categorical	type of job
marital	categorical	Few missing values along with information of marital status
education	categorical	Information about educational status
default	categorical	Answers the question if the person has credit in default
housing	categorical	Answers the question if the person has housing loan
loan	categorical	Answers the question if the person has personal loan
contact	categorical	contact communication type whether cellular phone or landline phone
month	categorical	last contact month of year
day of week	categorical	last contact day of the week
duration	Quantitative	last contact duration, in seconds
campaign	Quantitative	number of contacts performed during this campaign and for this client
pdays	Quantitative	number of days that passed by after the client was last contacted from a previous campaign.
previous	Quantitative	number of contacts performed before this campaign and for this client
poutcome	categorical	outcome of the previous marketing campaign
emp.var.rate	Quantitative	employment variation rate - quarterly indicator
cons.price.idx	Quantitative	consumer price index - monthly indicator
cons.conf.idx	Quantitative	consumer confidence index - monthly indicator
euribor3m	Quantitative	euribor 3 month rate - daily indicator
nr.employed	Quantitative	number of employees - quarterly indicator

After feature extraction random forest classifier and decision tree technique are used to train a model using original and subsequent data. Python is exclusively used for modelling all the datasets. ’sklearn’ package contains ’ensemble’ class where ’RandomForestClassifier’ function is available. This function is leveraged with parameters such as n_estimators, criterion, max_features, min_samples_split, min_samples_leaf. There are several ways and procedures for **knowledge distillation**, in this paper we are following a single approach for all the data set and compare the performance metrics of different models within the same data set. Random Forest algorithm by itself has the nature of ensembled decision trees and in this paper it is used in knowledge distillation process for all three data set and also it is used to compare with all other models that are processed with the same dataset. First step is to find the probability of classes in each observation of the data set. This is done by training a RandomForest Classifier with the features of the dataset using RandomForestClassifier function from sklearn.ensemble package. ’predict_proba’ is a useful function that predicts the probability for each classes using the model we trained with features of the target variable. Second step is to use the results as features on the next algorithm, that is probability obtained from previous step to train a Decision Tree classifier. Third step is to compare Decision Tree classifier on original data set and DecisionTree Classifier after knowledge distillation. Training a new model by considering the output of previous experiments were done on each data set and discussed further in the discussion section. The first data set is bi-classification data set, the outcome of the data set is to determine whether an adult belongs to either of the salary category mentioned in the target variable of the data set. The procedure is clearly explained in the figure-4. The first data set used in this paper is about adults living in the united states, the business problem is to find out the number of people who are earning more than 50K and less than or equal to 50K. A test data set is exclusively available for this problem, therefore splitting the data is not required for this data set. This dataset does not have class imbalance problem and Random forest is not affected by class imbalance problem.

Fig. 1. Different stages of experiments


Random Forest to Predict Probability
Training Decision Tree model with probability
Training a Decision Tree and Random Forest with original data set
Comparing accuracy and other metrics within data set
Applying other techniques and analysing using performance metric

The second data set used in this paper is about credit card defaulters. This data set does not have test data separately. Therefore certain percentage of original data set is selected at random and they are used for training. While the remaining observations are kept for evaluating the performance of the model that we have trained. Usually credit card defaulters data set tend to have class imbalance problems, this can be dealt with many techniques such as bootstrapping, sampling. In our experiment, it is not necessary to deal with class imbalance problems as we are not using techniques such as Logistic Regression, SVM or other classification techniques that tend to have more impact due to class imbalance issue. Random Forest and decision tree approach towards modelling is different from other classification techniques and it is not affected by class imbalance problem. The reason for this is Tree based approach will have conditions at each node and it takes the absolute value rather than estimating

Fig. 2. Correlation plot between features in table-1

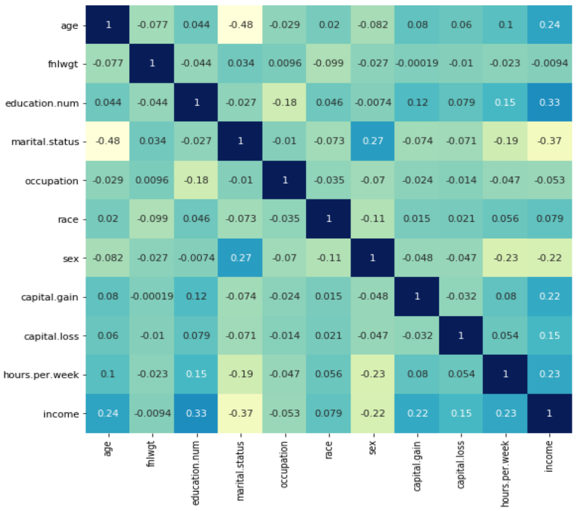
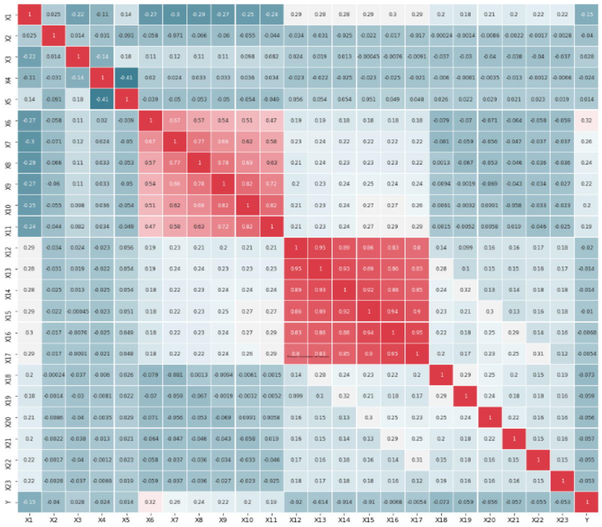


Fig. 3. Correlation plot between features in table-2



errors like other algorithm does.

The third data set used in this paper is bank marketing data set. This again has a bi-classification business problem. It is the analysis of the marketing strategy that involves whether a client or customer has purchased the plan after marketing the product to him. This business problem has a requirement to analyse the success factor of marketing and the corresponding business practice that has been followed. The target variable here is again categorical with ’yes’ and ’no’ being its classes. The above mentioned procedures are applied to the features of this data set in order to obtain best results and accuracy.

IV. Experiments

All the three dataset are processed separately as the needs for data cleaning and processing differs from each other. All the features in the dataset are checked for missing values and imputation is done for appropriate features. Certain records are discarded as imputation is not possible to replace the missing value. Correlation between the variables are plotted and the features are filtered, The correlation plot for the adult population dataset referenced at table-1 is given in figure-2. Correlation between the variables are plotted and the features are filtered based on the correlation and other feature selection methods, The correlation plot for the bank marketing dataset referenced at table-3 is given in figure-4. It is understood that correlation does not affect tree based models. Tree-based models have an innate feature of being robust to correlated features. When you drop a correlated variable to others, it will leave room for the tree to use one more variable in its trees[10]. Performance of the model could be affected since we are opening a new variable that is highly correlated. Although another variable is added, the importance of the correlated feature that has been removed is spread and stored among all other variables (It is specifically stored to the correlated features that has been removed). When highly correlated features are removed

Fig. 4. Correlation plot between features in table-3

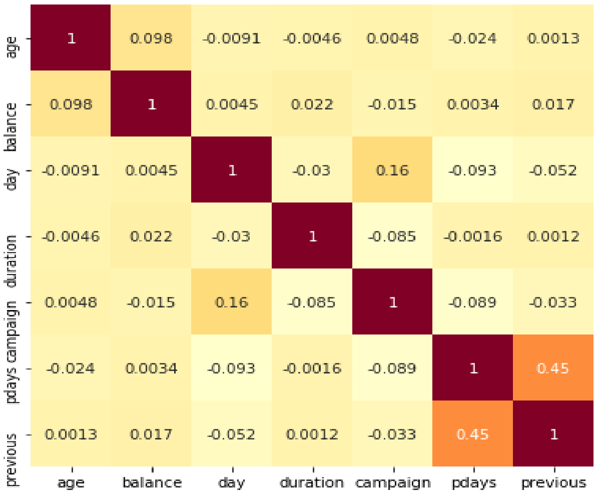


Fig. 5. Boxplot of features in dataset-1

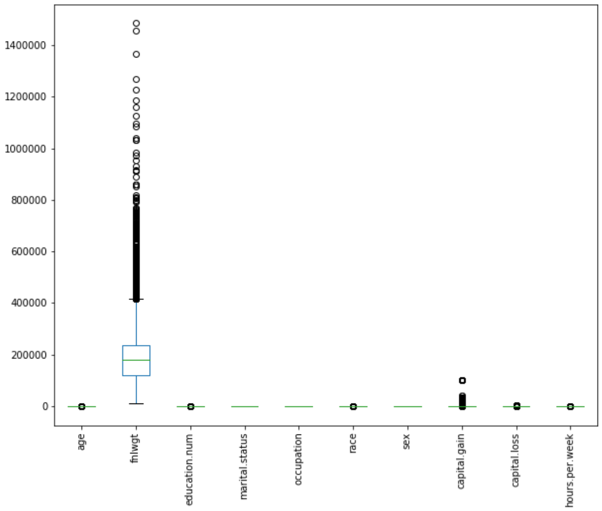
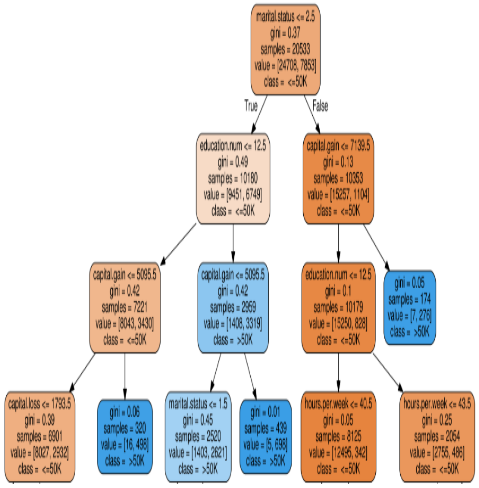


Fig. 6. Branch of RandomForest generated using python from Dataset-1



and if a model is trained without correlated features, the performance remains unchanged. This has been checked with two of the three data sets that are used in this paper. A branch from RandomForest classifier that was used to train with dataset-1 is added as an image in this paper as figure - 6. The figure can also be seen in the python notebook along with the code used to generate it.

Normalisation of data typically means to transform the observations 'x' into a function of 'x' and can be represented by 'f(x)' (where function is measurable, typically continuous function) such that they look normally distributed. Normalising the data is done so that all the input variables are treated the same way in the model and the coefficients of the model are not affected with respect to the units of the inputs. Although these are the benefits of normalising the data, in the case of tree based models normalising has no effect. Performance remains the same with/without normalisation of the data in all cases. This is because only the absolute values are considered in case of tree based approach and by default unlike other machine learning models errors are not included in the modelling technique.

One of the major advantages of random forests is their readability; in order to decrease the error, we can measure the importance of every feature. Random Forest technique are sometimes used in feature selection because of this nature. One way of removing feature significance is by permuting a given feature randomly and observing the changes in classification / regression. Another way of tracking the decrease in impurity or mid-square error is to be attributed to each feature for classification or regression as data flows through the trees in the forest.

The performance results of the knowledge distillation are tabulated from table IV to table X. The results are discussed in a detailed way in the discussion section.

V. Discussions

Evaluation of the models can be done using multiple metrics. One of the possibility is to assess the performance of the classification models of the random forest is the Area Under the Curve criterion. In addition, we may measure the performance of random forests against Area Under the Curve as a result of conventional logistic regression models where we use the same set of variables. The Area Under the Curve measure is computed on a range of comparisons between the forecast status of the event and the true status in respect of that event, taking into account all possible

TABLE IV

Table of Accuracy & Error Rate of Dataset - 1

Model Name	Accuracy	Error_Rate
RandomForest_gini(RF_G)	85.41%	14.61%
RandomForest_entropy(RF_E)	85.28%	14.72%
DecisionTree(DT)	80.46%	19.54%
DT_RF_G Distillation	84.09%	15.91%
DT_RF_E Distillation	83.67%	16.33%
DT_RF_G oneClass Distillation	79.53%	20.47%

TABLE V

Table of Accuracy & Error Rate of Dataset - 2

Model Name	Accuracy	Error_Rate
RandomForest_gini(RF_G)	82.48%	17.52%
RandomForest_entropy(RF_E)	82.52%	17.48%
DecisionTree(DT)	72.52%	19.54%
DT_RF_G Distillation	74.57%	27.48%

TABLE VI

Table of Accuracy & Error Rate of Dataset - 3

Model Name	Accuracy	Error_Rate
RandomForest_gini(RF_G)	89.79%	10.21%
RandomForest_entropy(RF_E)	89.94%	10.06%
DecisionTree(DT)	87.63%	12.37%
DT_RF_G Distillation	88.81%	11.19%

TABLE VII

Table of Precision & Recall & f_measure of Dataset - 1

Model Name	Precision	Recall	f_measure
RandomForest_gini(RF_G)	79.01%	52.08%	62.78%
RandomForest_entropy(RF_E)	78.92%	51.69%	62.47%
DecisionTree(DT)	58.59%	58.94%	58.77%
DT_RF_G Distillation	69.97%	57.2%	62.95%
DT_RF_E Distillation	71.41%	51.51%	0.5985%
DT_RF_G oneClass Distillation	56.41%	58.66%	57.51%

TABLE VIII

Table of Precision & Recall & f_measure of Dataset - 2

Model Name	Precision	Recall	f_measure
RandomForest_gini(RF_G)	73.41%	33.63%	46.13%
RandomForest_entropy(RF_E)	73.42%	33.86%	46.34%
DecisionTree(DT)	38.59 %	39.78%	39.17%
DT_RF_G Distillation	42.41%	40.0%	41.17%

TABLE IX

Table of Precision & Recall & f_measure of Dataset - 3

Model Name	Precision	Recall	f_measure
RandomForest_gini(RF_G)	61.14%	35.47%	44.90%
RandomForest_entropy(RF_E)	61.26%	38.49%	47.28%
DecisionTree(DT)	47.34%	47.23%	47.45%
DT_RF_G Distillation	50.05%	52.48%	47.83%

TABLE X

TABLE OF SENSITIVITY& SPECIFICITY OF DATASET - 1

Model Name	Sensitivity	Specificity
RandomForest_gini(RF_G)	52.08%	95.72%
RandomForest_entropy(RF_E)	51.69 %	95.73%
DecisionTree(DT)	58.94%	87.12%
DT_RF_G Distillation	57.2%	92.41%
DT_RF_E Distillation	51.51%	93.62%
DT_RF_G oneClass Distillation	58.66%	85.98%

TABLE XI

TABLE OF SENSITIVITY& SPECIFICITY OF DATASET - 2

Model Name	Sensitivity	Specificity
RandomForest_gini(RF_G)	33.63%	96.50%
RandomForest_entropy(RF_E)	33.86 %	96.48%
DecisionTree(DT)	39.78 %	81.89%
DT_RF_G Distillation	40.0 %	84.46%

TABLE XII

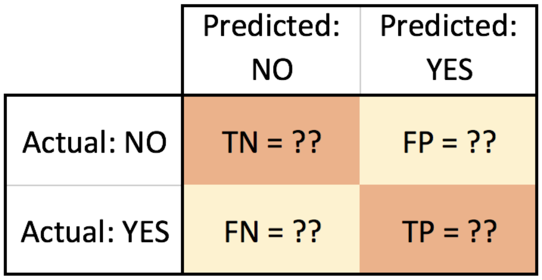
TABLE OF SENSITIVITY& SPECIFICITY OF DATASET - 3

Model Name	Sensitivity	Specificity
RandomForest_gini(RF_G)	35.47%	97.01%
RandomForest_entropy(RF_E)	38.49%	96.77%
DecisionTree(DT)	47.45%	92.96%
DT_RF_G Distillation	47.83%	94.25%

cut - off levels for the forecast values. In particular, the sensitivity that is the number of true positives with the total number of events and the specificity which is the number of true negatives with the total number of non - events of the confusion matrix are considered and summarised in a two - dimensional graph, resulting in a Receiver Operating Characteristic curve. We cannot use the Area Under the Curve evaluator with regards to the linear dependent variable, profit evolution, since both predicted and real values have more than two values. The evolution of profit represents the change in the profitability of the customer during the observed analysis window and can therefore have a wide range of predicted values. Another evaluation metric is confusion matrix, A clean and unambiguous way to present the prediction results of a classifier is to use a confusion matrix[2]. It has a another name contingency table[2].

Based on confusion matrix there are several other metrics that are used to calculate the performance of the model. It is clearly visible from the tables in experiment section that knowledge distillation improves the performance of the model. Although there is no significant improvement in the score, acceptable range of improvement is shown by the ensemble models when knowledge is transferred from parent to child. It also depends on dataset and **criterion** we choose for random forest. As entropy criteria shows better result in few instances and Gini index shows better in few other instances, it is highly difficult to conclude which criterion is better for knowledge distillation. On the other hand Random forest by itself has many properties and it can also act as a feature selector. Though feature selection is not required, it is tried in order to check the performance and accuracy of the model. Figure shows a box plot and the feature 'fnlwgt' shows more number of outlier observation. Therefore, it was normalised and used initially, no change in the results. Removed the variable and trained a model, no changes in the result. Random Forest proved to resilient and insensitive towards outliers, normalisation.

Fig. 7. Model of Confusion Matrix



	Predicted: NO	Predicted: YES
Actual: NO	TN = ??	FP = ??
Actual: YES	FN = ??	TP = ??

The dataset in this paper are small therefore no significant yet acceptable improvement is seen. Knowledge distillation might work well with large amount of data as more yet useful information will be fed to the model. The improvement shown by decision tree is through the information gained by Random forest with the original dataset. Passing probability is one of the many ways of doing knowledge distillation. If we analyse each table, every table shows some improvement in their percentage when knowledge distillation is carried out but the performance tend to decrease or show no improvement if features of the previous model are included in the dataset. The last observation in every dataset-1 table is the one class distillation that I tried to experiment but the accuracy rate and the performance of the model falls if we model with only one class. Either of the class the results are same. **Surprisingly, for the one class distillation model the results are lower than the original dataset's accuracy**. Although there is a class imbalance in few of the datasets, tree based models naturally takes care of this, therefore no steps were taken to solve the class imbalance problem.

Current knowledge distillation methods require full training data to distill knowledge from a large "teacher" network to a compact "student" network by matching certain statistics between "teacher" and "student" such as softmax outputs and feature responses[13]. This is not only time-consuming but also inconsistent with human cognition in which children can learn knowledge from adults with few examples[13]. Confusion matrix for a binary classification problem a table must have two rows and two columns[2]. The top of the table is the observed with class labels and down the sides are the predicted class labels[2]. Each cell contains the number of predictions made by the classifier that fall into that cell[2]. The predictions that are incorrect will clearly broke down into the two other cells. False Negatives are recurrence that the classifier has marked as no recurrence. False Positives are no recurrence that the classifier has marked as recurrence[2].

Precision is the number of TP(True Positive) divided by the number of TP(True Positive)[2]. In other words, the number of positive forecasts is divided by the total number of predicted positive class values. Positive predictive value is another name for Precision[2].

$$Precision = \frac{tp}{tp + fp} \tag{1}$$

Recall is the number of True Positives divided by the number of True Positives and the number of False Negatives[2]. To be precise the number of positive predictions divided by the number of positive class values in the test data[2]. Recall is also called Sensitivity or Real rate.

$$Recall = \frac{tp}{tp + fn} \tag{2}$$

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{3}$$

The F1 Score is another metric through which we calculate the performance of the model created. The formula to calculate this is as follows,

$$2 * ((precision * recall)/(precision + recall)) \tag{4}$$

It is also known as F Score or F Measure. The F1 score conveys the balance between the precision and the recall[2].

- The F1 for the All No Recurrence model is 2*((0*0)/0+0) or 0 [2].
- The F1 for the All Recurrence model is 2*((0.3*1)/0.3+1) or 0.46.
- The F1 for the CART model is 2*((0.43*0.12)/0.43+0.12) or 0.19.

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{5}$$

To select a model based on a balance between precision and recall, the F1 measure suggests that all Recurrence model is the one to beat and that CART model is not yet sufficiently competitive[2].
Following the brief discussion of evaluation metrics

VI. CONCLUSIONS

In all the three dataset mentioned in this paper knowledge distillation has improved the model by a smaller range. This may be due to the smaller size of the data. Random forest is one of the versatile and efficient algorithm among all the machine learning algorithms. It can be used right from selecting features for modelling to directly model a dataset. It is proven to be one of the best predictive algorithm till date. Knowledge distillation has proven to be one of the important methods through which a model's accuracy can be improved by a significant range. Although knowledge distillation is very useful in the neural networks it is highly difficult to identify and model a layer that transfers knowledge from one layer to another. [14]Deep Neural Networks have achieved huge success at a wide spectrum of applications from language modeling, computer vision to speech recognition[14]. However, nowadays, good performance alone is not sufficient to satisfy the needs of practical deployment where interpretability is demanded for cases involving ethics and mission critical applications[14]. Knowledge Distillation technique is also useful to attain good performance and interpretability in the neural network. Random Forest distillation to decision tree can be efficient for smaller dataset that requires complex modelling. In this work we have seen about the dataset that we used for modelling Random Forest classifier.Feature selection or feature extraction technique such as PCA is not used to choose features as random forest has the endancy to choose the best features for modelling. Correlation between the features are explained as a process of feature selection. Considered evaluation metrics to evaluate the machine learning model with ensemble machine learning models. In the dataset example which I have taken, the accuracy was misleading[2]. Sometimes it will be highly desirable to select a model that has lower accuracy because it will have a greater predictive power on the problem which we are facing[2]. In our performance metrics we used confusion matrix to avoid the accuracy paradox. Usage of confusion matrix confirms that the decision tree model shows improvement after knowledge distillation.

REFERENCES

[1] Amazon. *Simplify machine learning with XGBoost and Amazon SageMaker.* 2018. URL: https://aws.amazon.com/blogs/machine-learning/simplify-machine-learning-with-xgboost-and-amazon-sagemaker/.

[2] Jason Brownlee. *Classification Accuracy is Not Enough: More Performance Measures You Can Use.* 2014. URL: https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/.

[3] Mona Derakhshan. *Predicting credit card default.* 2016. URL: https://credit-default-predictor.herokuapp.com/.

[4] TG Dietterich. "Ensemble methods in machine learning". In: Jan. 2000, pp. 1–15. ISBN: 3-540-67704-6.

[5] Usama M Fayyd, Gregory P Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery: an overview". In: (1996).

[6] Georg Hermanutz. *Software using random forest for risk prediction of heart valve surgery patients.* 2017. URL: https://theses.cz/id/nq957c/thesis_georg_hermanutz.pdf.

[7] Ron Kohavi. *Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid.* 1996. URL: http://robotics.stanford.edu/~ronnyk/nbtree.pdf.

[8] Ron Kohavi. “Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid.” In: *Kdd*. Vol. 96. Citeseer. 1996, pp. 202–207.

[9] Philip Kotler and Kevin Lane Keller. *A framework for marketing management*. Prentice Hall, 2011.

[10] Laurae. *Ensembles of tree-based models: why correlated features do not trip them and why NA matters*. 2016. URL: `https://medium.com/data-design/ensembles-of-tree-based-models-why-correlated-features-do-not-trip-them-and-why-na-matters-7658f4752e1b`.

[11] Kevin Lemagnen. *Getting started with XGBoost.* 2018. URL: `https://cambridgespark.com/getting-started-with-xgboost/`.

[12] Susan Li. *Building A Logistic Regression in Python, Step by Step*. 2017. URL: `https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8`.

[13] Tianhong Li et al. “Knowledge Distillation from Few Samples”. In: *arXiv preprint arXiv:1812.01839* (2018).

[14] Xuan Liu, Xiaoguang Wang, and Stan Matwin. “Improving the Interpretability of Deep Neural Networks with Knowledge Distillation”. In: *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE. 2018, pp. 905–912.

[15] I Dan Melamed, Ryan Green, and Joseph P Turian. “Precision and recall of machine translation”. In: *Companion Volume of the Proceedings of HLT-NAACL 2003-Short Papers*. 2003.

[16] Sérgio Moro, Paulo Cortez, and Paulo Rita. “A data-driven approach to predict the success of bank telemarketing”. In: *Decision Support Systems* 62 (2014), pp. 22–31.

[17] mrisek. *Default-of-credit-card-clients.* 2016. URL: `https://github.com/mrisek/Default-of-credit-card-clients`.

[18] Vinay R. Rao. *How data becomes knowledge, Part 1 From data to knowledge.* 2018. URL: `https://www.ibm.com/developerworks/library/ba-data-becomes-knowledge-1/index.html`.

[19] P. Cortez S. Moro and P. Rita. *UCI Machine Learning repository - Bank Marketing Data Set.* 2016. URL: `https://archive.ics.uci.edu/ml/datasets/bank+marketing`.

[20] Ian H Witten et al. “KEA: Practical Automated Keyphrase Extraction”. In: *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific*. IGI Global, 2005, pp. 129–152.

[21] I-Cheng Yeh. *UCI Machine Learning repository - default of credit card clients Data Set.* 2016. URL: `http://archive.ics.uci.edu/ml/datasets/default%2Bof%2Bcredit%2Bcard%2Bclients`.

[22] I-Cheng Yeh and Che-hui Lien. “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients”. In: *Expert Systems with Applications* 36.2 (2009), pp. 2473–2480.

[23] Jason Yosinski et al. “Understanding neural networks through deep visualization”. In: *arXiv preprint arXiv:1506.06579* (2015).