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

Jayakrishnadeva Suresh Balasundaram - 1802552

***Abstract*— Ensemble methods are learning algorithms that build a set of classifiers and then classify new data points by voting their predictions (weighted). The original ensemble method is Bayesian averaging, but more recent algorithms incorporates concepts such as output coding, bagging and boosting for error correction. 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 consists of an ensemble of decision trees and is non - parametric, i.e. the data points are the parameters of the model. It naturally incorporates the selection of features 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 exaplained 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 [18] 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[16]. 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 analyze large amounts of data to discover relevant patterns and rules[18]. 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

behavior scoring and benchmarking (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 tabluar format,

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

TABLE I

Adult Data Description

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

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].

TABLE II

Default of Credit card clients Dataset

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.

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.

IV. Experiments

All the three dataset must be processed and only the features that give us the best result must be taken for modelling. 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-1.

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

TABLE III

BANK MARKETING DATA SET

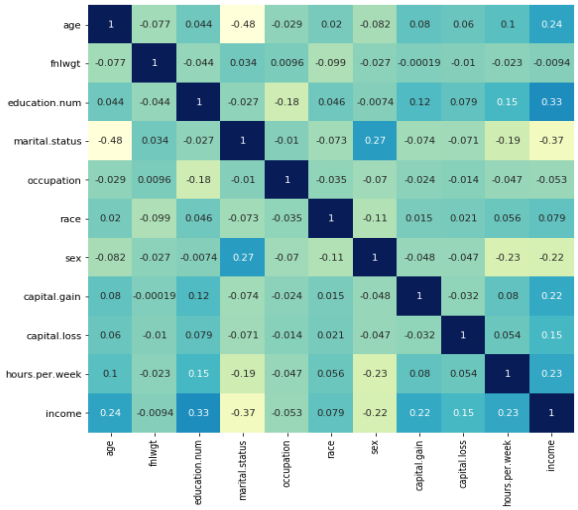

Fig. 1. Correlation plot between features in table-1

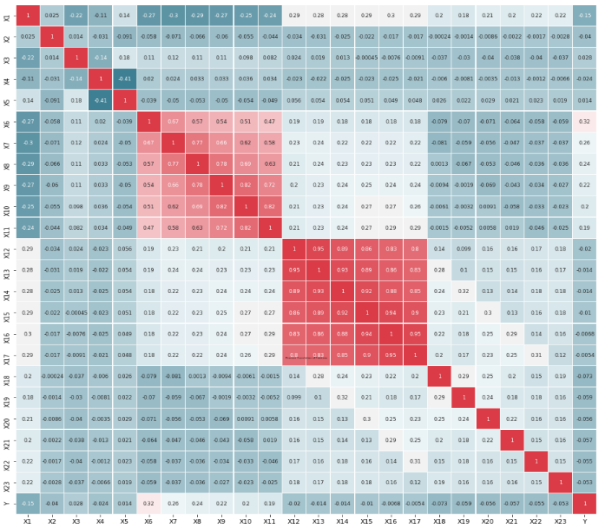

Fig. 2. Correlation plot between features in table-2

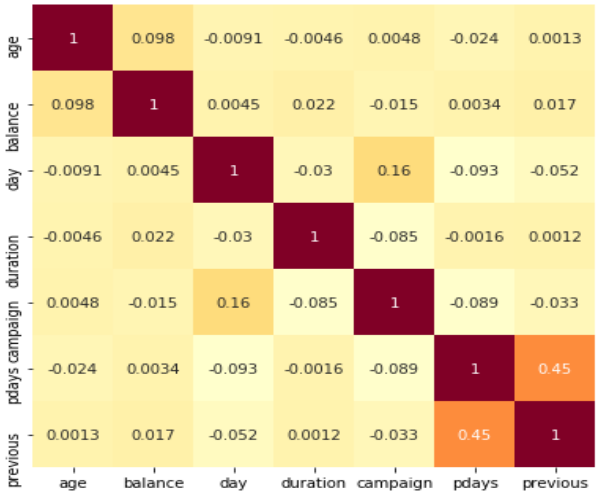

Fig. 3. Correlation plot between features in table-3

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-2.

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.

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

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

Fig. 4. Model of Confusion Matrix

forecast status of the event and the true status in respect of that event, taking into account all possible 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].

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].

VI. CONCLUSIONS

In this work we have seen about the dataset that we are going to use for modelling Random Forest classifier. Suggested feature selection or feature extraction such as PCA to choose features for best 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].
For example, in a problem where there is an imbalance between the class and if the imbalance is huge in number, then a model can predict the value for the majority class with very high predictions rate and this will lead to a high classification accuracy, the problem in that model is, it will not be useful in the actual problem domain[2].
This is called the Accuracy Paradox[2]. In our dataset, we might face similar problem. Therefore, accuracy must not be the sole reason to measure the model's performance and other metrics must be considered in order to train a effective model[2].

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

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.