A Comparative Analysis of five Machine Learning Algorithms for predicting Supply Chain profitability

Antonio Caruso
School of Computing
National College of Ireland
Dublin, Ireland
Email: x19203608@student.ncirl.ie

Abstract — Supply Chain Management has benefitted from advancement of Big Data Analytics and Machine Learning in predicting demand generation, anticipating customer service issues, automating tasks, mitigating risks and improving forecast accuracy [1]. The relevancy of this study is highlighted by the fact that in 2020, 95% of Supply Chain Planning vendors were estimated to incorporate Machine Learning into their solution due to the doubling amount of data every eighteen months[2]. This project aimed at understanding whether it was possible to use machine learning to predict orders profitability, by comparing the accuracy of five supervised machine learning algorithms on a Supply Chain Dataset of 180,519 values and identify which had highest accuracy for profitability prediction. The algorithms used were Decision Tree, Random Forest, K-Nearest Neighbour, Logistic Regression and Naïve Bayes. All the five classification algorithms were trained with 70% of the data with the remaining 30% used for testing the accuracy of each model, to prevent potential overfitting and unreliable accuracy results. The results obtained by application of confusion matrices showed K-Nearest Neighbour as the model that achieved the highest accuracy of 99.97% with a Kappa value of 1. This study took inspiration from previous work conducted on the dataset to bring new value.

Keywords—*Machine Learning Algorithms, Big Data Analytics, Supply Chain Management, Profitability Prediction, Decision Tree, Random Forest, K-Nearest Neighbour, Logistic Regression, Naïve Bayes,*

I. INTRODUCTION

A. Background on Supply Chain Management

Big data analytics and Machine Learning have the potential of being of tremendous impact in the field of supply chain management.

Despite research showing that Machine Learning has been applied on supply chain data with a specific focus on forecasting demand [3] and fraud prediction on transaction [4] , from a business perspective there is an apparent insufficient academic research on the impact of Machine Learning on this field. In fact, according to [5] automated models have not sufficiently influenced decision-making processes in the Supply Chain Risk Management (SCRM) field, and only 4% of studies have been conducted on large datasets.

The modern economy sees fast-paced market where orders, inventory and processes must be optimised to respond proactively to a global demand [6], therefore a general optimisation of the processes would theoretically determine an increase in profitability.

Fortunately, nowadays modern software technologies such as cloud computing and relational databases [7], can support discovery of insights and hidden patterns even amongst an unprecedented volume of data in an unimaginable way in the past.

This raised the question of how Machine Learning techniques can be effectively employed on large supply chain data containing numerous quantitative and qualitative information. In the following section, a brief background on these techniques and their functionality to this research is discussed.

B. Background of Machine learning for classification problems

Machine Learning is defined as an analytical technique built upon algorithms derived by computational statistics applied on data [8]. Again, [8] described how consistent labelled and relatable data would favour the application of supervised models for prediction, guaranteeing learning inference in the context of the data which in different scenarios would not be enough. Of a different opinion is [9], who claims how unsupervised models can produce faster results and are easier to implement than supervised models at the same cost, despite the latter presenting theoretical informational input that should ensure more effective results.

In the context of this research and the problems and objectives formulated in the following section, only supervised models in the form of classification algorithms for profitability prediction were examined, due to the high quantity of labelled numerical values associated with the orders in the supply chain dataset used, which suited perfectly this type of study for an accurate model. Moreover, this approach was employed as an attempt to conduct an in-depth study of different methodologies compared to [10], whose semi-supervised approach on the dataset employed for predicting transaction frauds with a reduced number of models and predictors, provided room for different approaches and research improvements.

C. Project goal and Research Objectives

The goal of this project was to perform a data analysis on the supply chain dataset to understand whether future orders would be profitable. Businesses can certainly harness the power of data mining techniques to existing data related to predict sales growth and product demand which directly affect profitability [11] .

A series of objectives were set in the beginning and accomplished at the end of this research;

- Conduct an exploratory data analysis to identify sales patterns, link or trends related to orders to gain useful insight for the research.
- Apply the best feature selection method to develop the classification models
- Compare the applied machine learning algorithms and identify which model had the highest accuracy in predicting whether orders will be profitable.

D. Problem Definition and Project Scope

The problem was formulated around the following research question:

Is it possible to use machine learning to predict orders profitability?

To answer the research question and achieve the pre-determined objectives, a large Supply Chain dataset was used for the analysis [1]. It covered areas of provisioning, production, sales, and commercial distribution of a various range of products. The dataset contained 180,249 rows and 53 variables, of which 22 were categorical and 31 numerical, with orders profitability as target variable. An initial exploratory analysis was deemed to provide relevant information about the data – e.g. information about orders, delivery, type of payments – before applying a feature engineering technique [2] to perform a reliable selection of the features as predictor variables to improve the power of the machine learning algorithms.

The project scope was determined around the development a classification model whereby the output variable would belong to two classes of “profitable” and “not profitable”, making it a binary classification problem [3]. This approach fell under the Supervised machine learning where the target variable was already labelled and known. When compared, each of the five algorithms provided an accuracy intended as probability of confidence of determined features to belong to each class.

II. ENGINEERING APPROACH

This section examines the advantages and disadvantages of classification algorithms applied before classification to compare their accuracy in predicting orders profitability.

A. Decision Tree:

The advantage of this method is that it can easily work with numerical and categorical data, resulting useful having both categories in the dataset chosen. Although a method that classifies data rapidly can be valuable, in the construction of the trees a small variation of the data can negatively impact the model [12].

B. Random Forest:

Random forest is a useful model for improving the performance of the decision tree. From the research conducted by [13], random forest appeared to be more suitable for a bigger dataset due to the random selection of features that improve the prediction power, unlike decision tree algorithms which could easily cause the model to overfit.

C. K-Nearest Neighbour:

It is an advantageous and suitable to large datasets algorithm for its functionality in nonparametric pattern classifications. Since it relies on one parameter K, despite its simplicity to implement the analysis, its application could incur in some problems such as imprecise testing samples and sensitivity to distance metric [14].

D. Logistic Regression:

One of the advantages of this model is to be one of the most intuiting and useful algorithms, especially useful for understanding the influence of numerous independent variables over on a single outcome variable. It could be useful for the dataset in place which present several numerical variables related to orders. Nevertheless, if data are highly discriminated towards certain variables, there is a risk of obtaining an unfair accuracy result that could impact the final comparison[15]

E. Naïve Bayes

Finally, Naïve Bayes is considered advantageous for performing with numerical variables when normal distribution is assumed. Nevertheless, the main disadvantage is that it works on unlikely assumption of independent predictors, therefore could be considered a bad estimator [16].

III. RELATED WORK

A. Machine learning to solve binary classification problems

Predicting whether orders were profitable falls under the broad application of Supervised Machine Learning. Knowing the target variable, the proposed techniques allow to build a model to solve a binary Classification problem based on predictor features, that can be applied in medical, finance and technology fields among many others [17]. Linear programming models have been proposed to classify whether an email would be spam, whether transactions could be considered fraudulent, or whether clients would respond to a promotional offer. The researcher [18] claimed that these models could generate biased results subject to dominant population groups of tested features especially for imbalanced datasets. By testing Minimised sum of deviations by proportion (MSDP) on a road casualties' dataset, he proved how more accurate results could be obtained in regard to mortal accidents depending on existing seat belt law, by adding the size of individual groups into the objective functions.

Being the chosen dataset balanced, complex models like MDSP were excluded since the initial step of this project. One of the models considered ideal for a classification problem is the Naïve Bayes model applied for an image recognition problem investigation[19] in which was claimed how this model would be accurate only when the assumptions that attributes to describe a target variables are independent is confirmed,

otherwise alternative models should be applied to avoid overfitting. A comparison of models was conducted by [20] for a water classification research that resulted in a 98.5% accuracy shown by a decision tree algorithm. The choice of factors directly related to cleanliness of the water were selected, but further research on selecting ideal parameters for improving models were suggested. Feature engineering as previously mentioned is fundamental to test the algorithms and can be conducted through different techniques. One of the most used is Correlation Based Feature Selection (CFS), which application determined the highest accuracy of 98.5% in a breast cancer research through application of K-nearest neighbour imputation method [21]. Despite accuracy being a common evaluation method for classification algorithms, for imbalanced data it could not be sufficient for model evaluation [22]. In fact, the consideration of a different method from the Accuracy was employed by [23] on two different datasets for classifying credit scoring. A comparison between Support vector machines, nearest neighbours, linear models and several more was determined by the measurement of the Area Under the operating characteristic curve (AUC). Despite being less used, [24] showed in their research a better validity compared to the accuracy model in application related to probability of the prediction, such as in ranking customers for a marketing analysis. However, not being those the purpose of the research, accuracy was chosen as method for evaluating the performance of the algorithms.

A. Machine Learning in Supply Chain industry

Investment in Machine learning can help supply chain businesses enhance product development, automate tasks, mitigate risks but above all improve forecast accuracy [1].

Research on Supply Chain Industry has been conducted by [25] who proposed a comparison of Distributed Random Forest and Gradient Boosting Machine learning, with a limitation of biased, redundant and missing data which determined a 20% of performance model increase for the Random Forest, but only when training ranged data.

In terms of comparing models for supply chain data, [26] did not come to a clear conclusion when comparing Decision Trees and SVM for an aerospace manufacturing supply chain dataset. This was due to the fact that in presence of imbalanced datasets and a wide range of metrics, interpretability should be prioritised over performance.

The chosen dataset for this study, was previously analysed by [27], and it is worth mentioning it for its contribution to the literature in the context of fraud prediction for online transactions. A semi-supervised model was applied by the author for the incompatibility of supervised models regarding misclassification to fraud detection and bias allocation to confusion matrices.

Predicting Orders profitability by applying supervised models, seemed to be an interesting challenge for this project and highly differentiated from fraud detection in seeking value in the context of supply chain data. Thus, this approach to predict order profitability with comparison of classification models can be considered as an attempt to learn from existing research and attempt a different technical approach. Also, gaining new insights and information that could contribute to improve forecast accuracy of profitability could be of exponential value to the entire industry.

IV. RESEARCH METHODOLOGY

The CRISP-Dm Process model was used to conduct the data mining process [28]. R programming language in Rstudio development environment was used to conduct the research, and visualisation graphs and charts were also implemented with the software Tableau.

A. Data Understanding

The large Supply Chain dataset was used for the analysis [29], with The DataExplorer package [30] applied to visualise the data structure and the attributes.

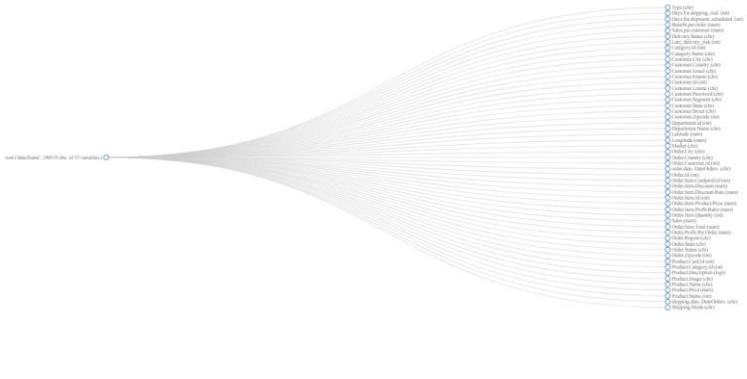

Figure 1 Data Structure (See appendix for full image)

The dataset contained initially 180,249 rows and 53 variables, of which 22 were categorical and 31 numerical. The target variable was identified in the "Order Profit per order", composed of positive numerical integers referring to profitable orders and negative integers referring to non-profitable orders.

B. Data Pre-processing

The data pre-processing allowed an initial analysis of the categorical and numerical variables in the dataset. After being loaded into Rstudio, the file was read with a read.csv function. A dplyr function [31] was used to remove duplicate or irrelevant columns, such as the ones related to customers' personal information from the orders placed. The only N/A missing values were present in the attributes related to product descriptions and zip codes. Before applying the planned algorithms, an exploratory analysis was conducted to explore trends and patterns about the data.

C. Exploratory Analysis

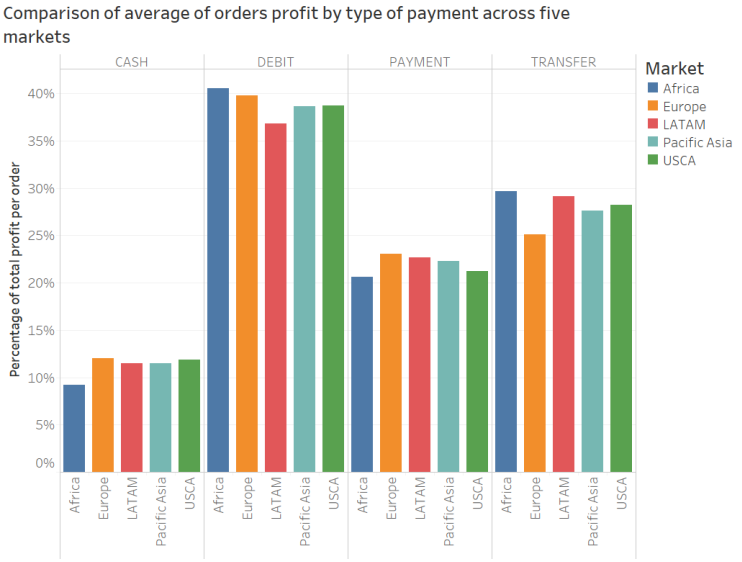

Figure 2 Bar chart: Order profit by type of payment (See appendix for full image)

An initial factual analysis on the main reputed attributes was executed to obtain an initial understanding of the data with Tableau for clear data visualisation. One of the first attributes analysed was the "type of payments" shown in the above chart to understand the contribution of type of payment to average orders profitability. The bar chart displays how orders paid by direct debit and transfers contributed the highest to orders profit, with the first reaching over 35% of profits with a 40% peak in , and the second showing a top of 30% profit with Africa and LATAM again being the top markets. European orders contributed mostly to Payment and Cash orders with respectively 23% and 12% of total orders profits.

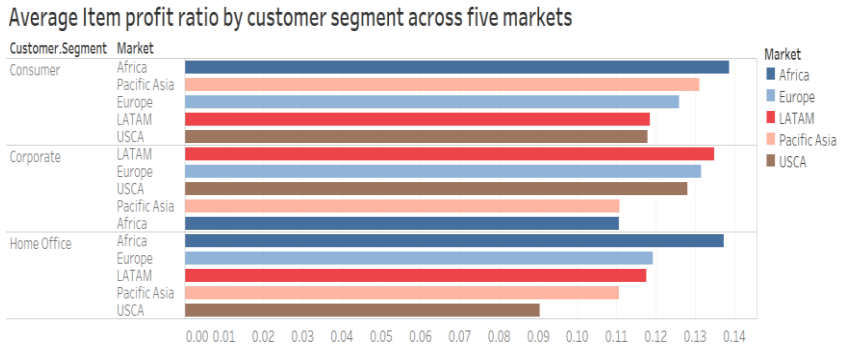

Figure 3 Bar Chart: Average item profit ratio by customer segment (see appendix for full image)

The average profit ratio by customer segment was also investigated across five markets. Items placed by consumers are worth between 12% and 14% profit, followed by Corporate segment with a range of 11% and 13.5%, and home office customers with a range 9% and 14%. Africa is the market where items sold to consumers and home office segments are profiting the highest, whereas the Market with the lowest item profit average is USCA for the home office segment with 9% profit.

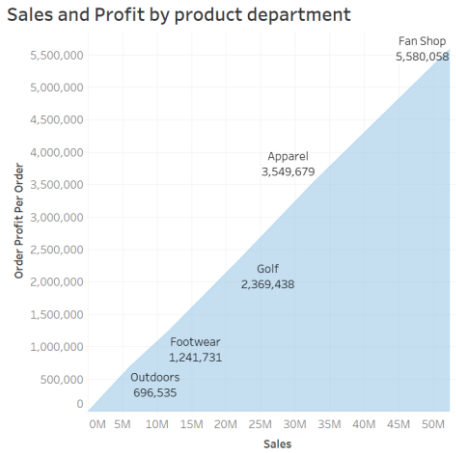

Figure 4 Area chart: Sales and Profit by product department (See appendix for full image)

The area chart above shows the sales and profit values generated by product departments, with a range between 500,000 and 5,500,000 of products that generated the highest profit. The fan shop products generated over 50 million of revenue for a profit of 5,500,000 million. The second category is apparel, which accounts for 30 million dollars in sales and 3,549,679 of profit. Golf and Footwear generated sales between 28 and 15 million and respectively a profit or almost 2.5 million and 1.5 million. The last notable product department is outdoors which generated slightly over 5 million in sales and a profit of almost 700,000.

D. Feature Engineering

The features analysed during the exploratory analysis were arbitrarily selected initially for their potential impact on orders profitability, and thus useful for the classification analysis. Despite providing interesting facts about the data, none of the trends and insights gathered resulted pivotal in conducting the final analysis.

The following step in fact entailed the choice of a Feature Engineering technique to select the ideal features as predictors to classify the target variable of "order profitability". It was fundamental to eliminate potential redundant features, ensure proper computability of the data, and prevent model overfitting. Principal Component Analysis (PCA) was chosen as a method for feature selection [32] that allows the spread of data in the form of a straight line displaying the highest variance in the data, to obtain uncorrelated principal components over the same samples. Since PCA works only with numerical variables, categorical variables were excluded, reducing the dataset to 14 numerical variables and 180,519 observations. With an explained variance of 31.4%, six variables whose most data points were closed to an explained variance close to 1 were identified in "Order.Item.Profit.Ratio", "Late_delivery_risk", "order.item quantity", "Order.Item.Discount.Rate", "Days.for.shipment..scheduled." and "Days.for.shipping.real". These variables were subsequently tested as predictors for the models.

V. MODELLING

The first step prior to the applications of models was to convert the "order.profit.per.order" target variable into a binary 0-1 variables, with 0 corresponding to non-profitable orders and 1 corresponding to profitable orders. The 'ifelse' function was applied for this [33] and after converting the variable into a factor, the summary function showed that the actual 33,784 orders were not profitable, whereas 146,735 were profitable.

An initial attempt on applying the six predicting variables, resulted in overfitting models, which for Classification Tree and Random Forest resulted in an unusual 100% accuracy. After a thorough investigation, the variable "Order.Item.Profit.Ratio" was identified as the responsible for overfitting due to its high correlation with the dependent variable which biased the final results. A data frame was then created with the final predictors and target variables renamed as below for clarity;

"order.profit.per.order"	Profit (Target Variable)
"Late_delivery_risk"	LateDelivery
"order.item quantity"	itemquantity
"Order.Item.Discount.Rate"	Discount
"Days.for.shipment..scheduled."	Shipment_Scheduled
"Days.for.shipping.real".	Shipment_Real

Table 1 Data frame for models application

A. Classification Tree

Classification Tree was the first model applied. First the dataset was split into standard 70% training set and 30% test set, resulting in 123,363 observations for training and 54,156 for test set. The 'rpart' algorithm [34] was applied.

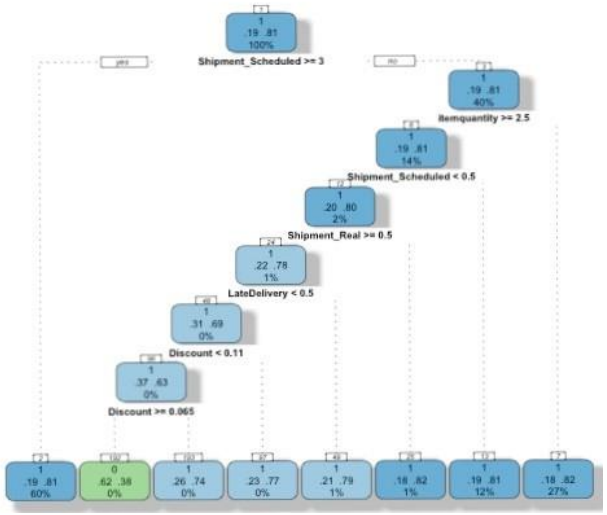

Figure 5 Classification Tree (See appendix for full image)

The tree displayed 8 terminal leaf nodes. The first node began the test with shipment_scheduled >= 3 days, splitting the prediction in 'Yes' and 'No', with the predicted value and the percentage of observation in the nodes. On the left side, 60% of orders observed classified as profitable, whereas on the right side of the tree, the classified 'No' for orders with shipment scheduled < 3, resulted in 40% of orders tested as profitable, of which 27% would have been profitable in case of item quantity per order < 2.5. The best Complexity parameter of 0 and the Gini index favouring larger partitions avoided tree overfitting [35].

The confusion matrix based on comparing the actual and predicted target column values in the test displayed the results as following:

CART	Pred: 0	Pred: 1
Actual: no	TN 1	FP 9,967
Actual: yes	FN 3	TP 44,185

Table 2 Confusion Matrix of the CART model

The output shows that of the 9,968 orders that actually were classified as non-profitable, the model predicted 1 as non-profitable (true negative) and 9,967 as profitable (false positive). Equally, of the 44,188 that were actually profitable, the model classified incorrectly 3 orders as non-profitable (false negative) and 44,185 (true positive) correctly as profitable.

CART	*Result*
Accuracy	*0.8159*
Kappa	*0.000001*

Table 3 Accuracy and Kappa results of the CART model

The accuracy of the Classification Tree resulted in 81.59% with Kappa as a metric ranging from -1 to +1 that compares observed accuracy with expected accuracy equalling to 0,000001 as an index of reliability, positively accepted as rater reliability should not be under a certain threshold only for health research [36].

B. Random Forest

The 'randomforest' algorithm [37] was applied for the second model with 70% sample of the dataset to build the model.

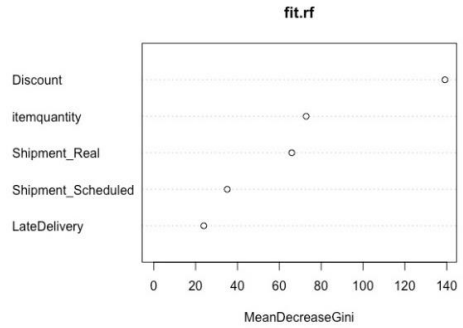

Figure 6 Mean Decrease Gini values of the Random Forest model

500 trees were created with two variables tried at each split was obtained with an OOB estimate error of 18.66%, meaning that the model has 82.34% out of sample accuracy. The mean Decrease Gini measured how each variable contributed to the homogeneity of the nodes, thus it can be considered more important for the evaluation. The graph shows the highest value for "Discount". By adding more trees compared to Classification Tree, potential overfitting was preventively excluded. Subsequently, the confusion matrix again produced the following results.

Random Forest	Pred: 0	Pred: 1
Actual: 0	TN 0	FP 102,205
Actual: 1	FN 0	TP 43,951

Table 4 Confusion Matrix of the Random Forest model

The output shows that of the 102,205 orders that were actually classified as non-profitable, the model predicted 0 as non-profitable (true negative) and 102,205 as profitable (false positive). Equally, of the 43,951 that were actually profitable, the model classified incorrectly 0 orders as non-profitable (false negative) and 43,951 (true positive) correctly as profitable.

Random Forest	*Result*
Accuracy	*0.8112*
Kappa	*0*

Table 5 Table 3 Accuracy and Kappa results of Random Forest model

The Random Forest model resulted in an accuracy of 81.12% and Kappa value of 0.

C. K-Nearest Neighbour:

The third model applied was the K-nearest neighbour. Firstly, data normalisation was employed to avoid biases. Subsequently training data were used to build the model and the test data for validation, again with 70% training set and 30% test set to build the model with the class package and KNN function [38]. In order to identify the optimum value for K, the square root of total number of observations of train dataset (123,363 observation) was calculated and resulted in 351.23. Nevertheless, the value was halved and the performance was evaluated on K= 175 because of too many ties in KNN which impeded to compute the algorithm on K= 352.

K-NN	Pred: 0	Pred: 1
Actual: 0	TN 10,126	FP 0
Actual: 1	FN 15	TP 44,015

Table 6 Confusion Matrix of the KNN model

The output from the confusion matrix shows that of the 10,126 orders that actually were classified as non-profitable, the model predicted 10,126 as non-profitable (true negative) and 0 as profitable (false positive). Equally, of the 44,030 that were actually profitable, the model classified incorrectly 15 orders as non-profitable (false negative) and 44,015 (true positive) correctly as profitable.

K-Nearest Neighbour	*Result*
Accuracy	*0.9997*
Kappa	*0.9991*

Table 7 Accuracy and Kappa results of the KNN model

The accuracy value of the K-Nearest Neighbour resulted in 99.97% with Kappa equalling to 0.9991.

D. Logistic Regression

Once the dataset was split again into 70% training set and 30% test data set, the 'glm' function [39] was applied to fit the regression model of type 'logit' with parameter family equal to binomial. Before running the confusion matrix, the anova function was run and showed how the best fit to the data was for the 'itemquantity' variable which had the lowest variance.

Logistic Regression	Pred: 1	Pred: 0
Actual: 1	TP 44,086	FN 10,070
Actual: 0	FP 0	TN 0

Table 8 Confusion Matrix of the Logistic Regression model

The output shows that, of the 54,156 orders classified as profitable, the model predicted 44,086 correctly as profitable (True Positive) and 10,070 incorrectly as Non profitable (False Negative).

Logistic Regression	*Result*
Accuracy	*0.8141*
Kappa	*0*

Table 9 Accuracy and Kappa results of the Logistic Regression model

The accuracy of the Logistic Regression model in predicting profitability resulted in 81.41% accuracy, with Kappa equalling to 0.

E. Naïve Bayes

The Naïve Bayes classifier is a simple probabilistic classifier based on the Bayes theorem but with strong assumptions regarding independence. It replaced the Support Vector Machine initially planned for its

computability issues encountered. Once split the dataset again into 70% training and 30% test datasets, the Caret package with nb algorithm [40]was used for easy 10-k cross validation and tuning. The train_control function was employed then to create the model before visualising the confusion matrix as following.

Naïve Bayes	Pred: 0	Pred: 1
Actual: 0	TN 0	FP 0
Actual: 1	FN 9,968	TP 44,188

Table 10 Confusion Matrix of the Naïve Bayes model

The output shows that of 54,156 orders actually classified as profitable, the model predicted 9,968 incorrectly as non-profitable (False Negative) and 44,188 as profitable (True Positive)

Naïve Bayes	*Result*
Accuracy	*0.8129*
Kappa	*0*

Table 11 Accuracy and Kappa results of Naïve Bayes model

The accuracy of the Naïve Bayes model resulted in 81.29% accuracy, with Kappa equalling to 0 as an index of reliability.

VI. COMPARISON OF RESULTS

	CART	***RF***	***KNN***	***Logistic Regression***	***NB***
Accuracy	81.59%	81.12	99.97%	81.41%	81.29%
Kappa	0.00001	0	0.9991	0	0

Table 12 Comparison table of Accuracy and Kappa results

The comparison table above shows the K-Nearest Neighbour is the model that performed best with 99.97% accuracy and Kappa close to 1 as indication of an almost perfect model for predicting orders profitability. All the other models resulted in the 81% range accuracy, therefore, no model except for KNN outperformed each other, and each of them could be considered highly valid, having an accuracy above 80% (considered of high standard in Machine Learning). The results satisfy the purpose of the research, following an initial attempt of analysis that included the variable "Order.Item.Profit.Ratio", that had produced improbable perfect results of 100% for CART and RF and above between 98% and 99% for the remaining models.

VII. CONCLUSIONS & FUTURE WORK

A. Conclusion

This study has been a great opportunity to investigate the power of several Machine Learning methods in an interesting field such as Supply Chain.

The results found allowed to respond to the research question that, by correctly testing and employing specific methods to perform the classification models, it was possible to predict orders profitability with a high degree of confidence. It can be stated also that the research objectives were met. In fact, albeit the exploratory analysis supported the final analysis to a certain degree, it helped better understand the data, identify potentially relevant variables, and remove redundant features that could have negatively compromised the results. Moreover, the chosen feature selection method successfully allowed to select the right predictors to obtain functional model results.

Testing and re-evaluating the models has been key to obtaining satisfactory results. For instance, Logistic Regression was excluded at mid-stage of the project for its lacking performance when the "Profit Ratio" variable was present. Support Vector Machine, initially considered, was excluded at the end in favour of Naïve Bayes for the hard-computational efficiency experienced once the biased "Profit Ratio" variable was excluded.

The study took inspiration from a research on the data set conducted by [10], whereby different variables where selected for transactional fraud predictions and semi-supervised models were applied. The approach taken for this project has, therefore, analysed a different problem from a different

perspective, demonstrating how with a different object of predictions for the chosen dataset, satisfactory accuracy results could be achieved.

B. Future Work

Future work could entail application on unsupervised models such as clustering, to discover hidden purchasing patterns, or to investigate further the influence of product categories on sales and type of payments across different countries or continents.

ACKNOWLEDGMENTS

This data analysis project was conducted towards the completion and conferment of the award of degree of a Higher Diploma in Science in Data Analytics pursued at the National College of Ireland. The author would like to thank John Bohan for his useful guidance and supervision provided throughout the semester, as well as other lecturers who provided with great professionalism the necessary knowledge to reach this achievement.

REFERENCES

[1] H. Canitz, "Machine Learning in Supply Chain Planning: When Art & Science Converge" in *Discovery Service for NCI Library*. 2019. [Online]. Available: http://eds.b.ebscohost.com/eds/detail/detail?vid=3&sid=cbd97cef-75be-4852-8480-9332802cf5e3%40pdc-v-sessmgr04&bdata=JkF1dGhUeXBlPWlwLGNvb2tpZSxzaGliJnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#AN=136112885&db=bsu. [Accessed Oct. 24, 2020].

[2] Supply Chain Management Review. "Making the case for AI and Machine Learning in Supply Chain Planning: How companies can ensure peak supply chain performance with an autonomous engine that continuously senses, analyzes, and updates demand planning parameters in real-time." in *Discovery Service for NCI Library*, 2020 [Online] Available at: http://eds.a.ebscohost.com/eds/detail/detail?vid=39&sid=32840d0a-d283-4002-bfc2-2ec54711b976%40sessionmgr4006&bdata=JkF1dGhUeXBlPWlwLGNvb2tpZSxzaGliJnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#AN=edsgcl.617150090&db=edsgbe [accessed Oct. 16, 2020].

[3] N. Vandeput "Machine Learning for Supply Chain Forecasting", *by Towards Data Science.*, 2019. [Online]. Available: https://towardsdatascience.com/machine-learning-for-supply-chain-forecast-66ef297f58f2 [Accessed Nov. 07, 2020].

[4] F. V. Constante-Nicolalde, P. Guerra-Terán, and J. L. Pérez-Medina, "Fraud Prediction in Smart Supply Chains Using Machine Learning Techniques," in *Communications in Computer and Information Science*, 2020. Available: https://www.researchgate.net/publication/339640750_Fraud_Prediction_in_Smart_Supply_Chains_Using_Machine_Learning_Techniques [Accessed Nov. 06, 2020]

[5] G. Baryannis, S. Validi, S. Dani, and G. Antoniou, "Supply chain risk management and artificial intelligence: state of the art and future research directions," *International Journal of Production Research*, 2019. [Online]. Available: http://eds.b.ebscohost.com/eds/detail/detail?vid=14&sid=dccc4edc-9639-4887-aad2-64d6ec13e17e%40pdc-v-sessmgr02&bdata=JkF1dGhUeXBlPWlwLGNvb2tpZSxzaGliJnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#AN=135800702&db=bsu [Accessed Nov. 06, 2020]

[6] M. Natarajarathinam, I. Capar, and A. Narayanan, "Managing supply chains in times of crisis: A review of literature and insights," *International Journal of Physical Distribution & Logistics Management*, 2009, [Online]. Available: http://eds.b.ebscohost.com/eds/detail/detail?vid=16&sid=dccc4edc-9639-4887-aad2-64d6ec13e17e%40pdc-v-sessmgr02&bdata=JkF1dGhUeXBlPWlwLGNvb2tpZSxzaGliJnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#AN=45002851&db=bsu. [Accessed Nov. 06, 2020]

[7] Y. H. Kuo and A. Kusiak, "From data to big data in production research: the past and future trends," *International Journal of Production Research*, 2019. [Online] Available: http://eds.b.ebscohost.com/eds/detail/detail?vid=18&sid=dccc4edc-9639-4887-aad2-64d6ec13e17e%40pdc-v-sessmgr02&bdata=JkF1dGhUeXBlPWlwLGNvb2tpZSxzaGliJnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#AN=137944657&db=bsu. [Accessed Nov. 06, 2020]

[8] D. Viberg and M. H. Eslami "The Effect of Machine Learning on Knowledge-Intensive R&D in the Technology" *in Discovery Service for NCI Library*, 2020. [Online] Available: http://eds.b.ebscohost.com/eds/detail/detail?vid=10&sid=6acd227e-c35e-451c-b64b-03a46e8529c5%40pdc-v-sessmgr01&bdata=JkF1dGhUeXBlPWlwLGNvb2tpZSxzaGliJnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#db=ent&AN=142604298. [Accessed Oct. 18, 2020].

[9] H. Qiao, X. Xin, L. David, "Supervised vs Unsupervised Models: A Holistic Look at Effort-Aware Just-in-Time Defect Prediction" *in IEEE Conference Publication*, 2019. [Online]. Available: https://ezproxy.ncirl.ie:2102/document/8094418?arnumber=8094418 [Accessed Oct. 18, 2020].

[10] F.V. Constante-Nicolalde, P. Guerra-Terán, J. L. Pérez-Medina, "Fraud Prediction in Smart Supply Chains using Machine Learning Techniques." *in Discovery Service for NCI Library*, 2020. [Online] Available: http://eds.b.ebscohost.com/eds/detail/detail?vid=46&sid=dccc4edc-9639-4887-aad2-64d6ec13e17e%40pdc-v-sessmgr02&bdata=JkF1dGhUeXBlPWlwLGNvb2tpZSxzaGliJnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#AN=edselc.2-52.0-85082389193&db=edselc [Accessed Oct. 23, 2020].

[11] Minderest. "Big Data: The Role of Predictive Analytics in Sales Growth.", 2020. [Online]. Available: https://martechseries.com/sales-marketing/sales-intelligence/the-role-of-predictive-analytics-in-sales-growth/ [Accessed Nov. 07, 2020).

[12] M. Somvanshi, P. Chavan, S. Tambade, and S. v. Shinde, "A review of machine learning techniques using decision tree and support vector machine" 2016, [Online]. Available: http://eds.b.ebscohost.com/eds/detail/detail?vid=26&sid=dccc4edc-9639-4887-aad2-64d6ec13e17e%40pdc-v-sessmgr02&bdata=JkF1dGhUeXBlPWlwLGNvb2tpZSxzaGliJnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#AN=edseee.7860040&db=edseee. [Accessed Nov. 06, 2020]

[13] J. Ali, R. Khan, N. Ahmad, I. Maqsood "Random Forests and Decision Trees." in *Research Gate*, 2012. [Online]. Available https://www.researchgate.net/publication/259235118_Random_Forests_and_Decision_Trees [Accessed Nov. 07, 2020].

[14] J. Wang, P. Neskovic, and L. N. Cooper, "Improving nearest neighbor rule with a simple adaptive distance measure," *in Pattern Recognition Letters*, 2007. [Online] Available: https://www.researchgate.net/publication/221161705_Improving_Nearest_Neighbor_Rule_with_a_Simple_Adaptive_Distance_Measure. [Accessed Nov. 05, 2020]

[15] S. Radovanović, A. Petrović, B. Delibašić, & M. Suknović, n.d. "Enforcing fairness in logistic regression algorithm." [Online] Available: https://ezproxy.ncirl.ie:2102/stamp/stamp.jsp?tp=&arnumber=9194676&tag=1 [Accessed Nov. 07, 2020].

[16] G. Chauhan "All about Naive Bayes" *in Towards Data Science*, 2018. [Online] Available: https://towardsdatascience.com/all-about-naive-bayes-8e13cef044cf

[17] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning: A review of classification and combining techniques," *in Artificial Intelligence Review*, 2006, [Online] Available: https://www.researchgate.net/publication/226525180_Machine_learning_A_review_of_classification_and_combining_techniques [Accessed Nov. 07, 2020].

[18] M. O. Olusola and S. I. Onyeagu, "On the binary classification problem in discriminant analysis using linear programming methods," *in Operations Research and Decisions*, 2020. [Online] Available: http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.desklight-0984e5e4-1bc0-4c37-8f4c-3a21d42f73cd [Accessed Oct. 18, 2020].

[19] H. Langseth and T. D. Nielsen, "Latent classification models for binary data," *in Pattern Recognition*, 2009. [Online] Available: https://www.sciencedirect.com/science/article/pii/S0031320309001708 [Accessed Oct. 18, 2020].

[20] N. Radhakrishnan and A. S. Pillai, "Comparison of Water Quality Classification Models using Machine Learning," *in Discovery Service for NCI Library*, 2020. [Online] Available: http://eds.b.ebscohost.com/eds/detail/detail?vid=31&sid=dccc4edc-9639-4887-aad2-64d6ec13e17e%40pdc-v-sessmgr02&bdata=JkF1dGhUeXBlPWlwLGNvb2tpZSxzaGliJnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#AN=edseee.9137903&db=edseee, [Accessed Oct. 25, 2020].

[21] M. Anitha, S. Gayathri, S. Nickolas, M.S. Bhanu, "Feature Engineering based Automatic Breast Cancer Prediction." *in Discovery Service for NCI Library*, 2020. [Online] Available: http://eds.b.ebscohost.com/eds/detail/detail?vid=3&sid=813d5340-eb27-43ca-8b41-417112cb80a5%40pdc-v-sessmgr04&bdata=JkF1dGhUeXBlPWlwLGNvb2tpZSxzaGliJnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#AN=edseee.9182855&db=edseee [Accessed Oct. 23, 2020].

[22] J. Brownlee. "Classification Accuracy is Not Enough: More Performance Measures You Can Use." *in Machine Learning Process.*, 2014, [Online] Available: https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/ [accessed Nov. 07, 2020].

[23] L. S. M. Junior, F. M. Nardini, C. Renso, and J. A. F. Macedo, "An empirical comparison of classification algorithms for imbalanced credit scoring datasets" in *Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA*. 2019, [Online] Available: http://eds.b.ebscohost.com/eds/detail/detail?vid=37&sid=dccc4edc-9639-4887-aad2-64d6ec13e17e%40pdc-v-sessmgr02&bdata=JkF1dGhUeXBlPWlwLGNvb2tpZSxzaGliJnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#AN=edseee.8999279&db=edseee. [Accessed Oct. 27, 2020].

[24] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, 2005. [Online] Available: http://eds.b.ebscohost.com/eds/detail/detail?vid=39&sid=dccc4edc-9639-4887-aad2-64d6ec13e17e%40pdc-v-sessmgr02&bdata=JkF1dGhUeXBlPWlwLGNvb2tpZSxzaGliJnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#AN=edseee.1388242&db=edseee. [Accessed Oct. 28, 2020].

[25] . S. Khodabandehlou and M. Zivari Rahman, "Comparison of supervised machine learning techniques for customer churn prediction based on analysis of customer behavior," *in Journal of Systems and Information Technology*, 2017. [Online] Available: http://eds.b.ebscohost.com/eds/detail/detail?vid=41&sid=dccc4edc-9639-4887-aad2-

64d6ec13e17e%40pdc-v-sessmgr02&bdata=JkF1dGhUeXBlPWlwLGNvb2tpZSxzaGliJnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#AN=edsemr.10.1108.JSIT.10.2016.0061&db=edsemr. [Accessed Oct. 28, 2020].

[26] S. Islam and S. H. Amin, "Prediction of probable backorder scenarios in the supply chain using Distributed Random Forest and Gradient Boosting Machine learning techniques," *in Journal of Big*, 2020. [Online] Available: http://eds.b.ebscohost.com/eds/detail/detail?vid=43&sid=dccc4edc-9639-4887-aad2-64d6ec13e17e%40pdc-v-sessmgr02&bdata=JkF1dGhUeXBlPWlwLGNvb2tpZSxzaGliJnNpdGU9ZWRzLWxpdmUmc2NvcGU9c2l0ZQ%3d%3d#AN=edsdoj.bec54b0a76c4072a29041ecc91b48b8&db=edsdoj. [Accessed Oct. 28, 2020].

[27] G. Baryannisa, S. Danib, G. Antonioua, "Predicting supply chain risks using machine learning: The trade-off between performance and interpretability." i*n Science Direct*, 2019. [Online] Available: https://www.sciencedirect.com/science/article/pii/S0167739X19308003. [Accessed Nov. 05, 2020].

[28] V. Kotu, B. Deshpande "Data Mining Process - an overview" *in ScienceDirect Topics*. 2015. [Online] Available: https://www.sciencedirect.com/topics/computer-science/data-mining-process [Accessed Oct. 25, 2020].

[29] S. Tiwari "DataCo: Smart supply chain for big data analysis" *in Kaggle*. 2019. [Online] Available: https://www.kaggle.com/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis?select=DataCoSupplyChainDataset.csv [Accessed Oct. 25, 2020].

[30] B. Cui "Automate Data Exploration and Treatment" *in RDocumentation* [Online] Available: https://www.rdocumentation.org/packages/DataExplorer/versions/0.8.2 [Accessed Nov. 25, 2020].

[31] DataScience Made Simple "Drop column in R using Dplyr - drop variables" [Online] Available: https://www.datasciencemadesimple.com/drop-variables-columns-r-using-dplyr/ [Accessed Nov. 07, 2020].

[32] L. Hayden "PCA Analysis in R" *in DataCamp.* 2018. [Online] Available: https://www.datacamp.com/community/tutorials/pca-analysis-r [Accessed Nov. 07, 2020].

[33] O. Smith "(Tutorial) IF ELSE Function in R" *in DataCamp.* 2020. [Online] Available: ." https://www.datacamp.com/community/tutorials/if-else-function-r?utm_source=adwords_ppc&utm_campaignid=898687156&utm_adgroupid=48947256715&utm_device=c&utm_keyword=&utm_matchtype=b&utm_network=g&utm_adpostion=&utm_creative=255798340456&utm_targetid=aud-299261629574:dsa-429603003980&utm_loc_interest_ms=&utm_loc_physical_ms=1007850&gclid=CjwKCAiAqJn9BRB0EiwAJ1SztZ8n3KIchCJSuzjnfaGG1IN5noeQk3ThBAtvrwtUV2Idu3exszJVZRoC6ZEQAvD_BwE (accessed Nov. 07, 2020).

[34] RDocumentation "Recursive Partitioning And Regression Trees" [Online] Available: https://www.rdocumentation.org/packages/rpart/versions/4.1-15/topics/rpart (accessed Nov. 07, 2020).

[35] Learn by Marketing "Data Mining + Marketing in Plain English" 2020 [Online] Available: http://www.learnbymarketing.com/481/decision-tree-flavors-gini-info-gain/ (accessed Nov. 27, 2020).

[36] M. L. McHugh "Interrater reliability: the kappa statistic" *in Biochem Med* (Zagreb) 2012 [Online] Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/#:~:text=Cohen's%20kappa,-Cohen's%20kappa%2C%20symbolized&text=Cohen%20suggested%20the%20Kappa%20result,1.00%20as%20almost%20perfect%20agreement (accessed Nov. 27, 2020).

[37] RDocumentation "Classification And Regression With Random Forest" [Online] Available: https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/randomForest (accessed Nov. 23, 2020).

[38] RDocumentation "K-Nearest Neighbour Classification" [Online] Available: https://www.rdocumentation.org/packages/class/versions/7.3-17/topics/knn (accessed Nov. 23, 2020).

[39] RDocumentation "Fitting Generalized Linear Models" [Online] Available: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/glm (accessed Nov. 23, 2020).

[40] RDocumentation "Naive Bayes Classifier" [Online] Available: https://www.rdocumentation.org/packages/e1071/versions/1.7-3/topics/naiveBayes (accessed Nov. 23, 2020).