ISYS3420: Machine learning for Decision-Makers

Assessment 1: Literature Review

Prepared By: Rokshana Pervin

Executive Summary:

Machine learning is all about teaching machines how to learn. In the real-world field, we can see three types of machine learning. These are: supervised, unsupervised, and semi-supervised machine learning. In this literature review, three types of machine learning's problems and its solutions have been discussed from the business area as well as three decisions have been taken using these solutions. Here, linear regression algorithm is used to minimise the supervised machine learning problem, isolation forest model or iForest algorithm is applied to solve unsupervised machine learning problem, and self-training method is used to mitigate the semi-supervised machine learning problem. To take those above decisions, some key points have been followed such as described about three different problems that use three different techniques to inform decision-making, thoroughly explained how machine learning was applied to business opportunities, the impact of the techniques in business opportunities, as well as insightfully demonstrated why machine learning used and its advantages in the business opportunities. Apart from this, insightfully analysed to evaluate all three solutions for solving business problems and considered some limitations by providing minimisations of these techniques so that stakeholders can aware regarding the decisions of their future business using machine learning solutions.

Introduction:

Nowadays, stakeholders are confronting with many business problems of machine learning in their business. These obstacles are making them behind to take appropriate decisions for their future business outcomes and goals. The purpose of this literature is to identify three business problems related to machine learning and apply three different machine learning techniques to enhance the stakeholders' business opportunities while making decisions regarding their businesses.

Evaluation Framework:

1st Business Problem:

In today's business, increasing revenue is a big challenge for every stakeholder. Some retail stores are struggling to overcome this challenge although they are confronting with many hindrances such as consumer attributes, behaviours as well as demographic information.

Technique: Supervised machine learning

In order to mitigate this above challenge, store managers are using some machine learning techniques where supervised machine learning solution is remarkable. Khushbu and Suniti (2018) explain that linear regression is one of the supervised machine learning techniques which has statistical procedure and prediction for dependent variable based on independent variables. In this process, a target variable or response variable finds its fitting line relies on its predictive variable. It is mainly statistical approach where dependent and independent variables are as shown in follows equation. Here, y is representing as a target or dependent variable and x is depicted as an independent variable as well as a and b illustrates as co-efficient (Scott 2021).

$$y = ax + b$$

In this case, it can anticipate that one grocery store where manager is investing money to make the online advertisement to extend store's publicity and to get appropriate revenue. The above equation can apply in this scenario where x will represent as an amount which is invested in online advertisement and y will represent as revenue for the grocery store (Scott 2021). As regression is prioritising it's best fitting line, x and y values will draw a straight line which will minimise the distance of the data points and aid to predict y values to understand the predictions of the distances how far staying from the actual values. Here, if the distance is described as high value from the other distance in the fitting line, it will be a residual or error. If the residuals distributed normally, it would show the assumption of the linear regression of x and y or the straight line (Scott 2021).

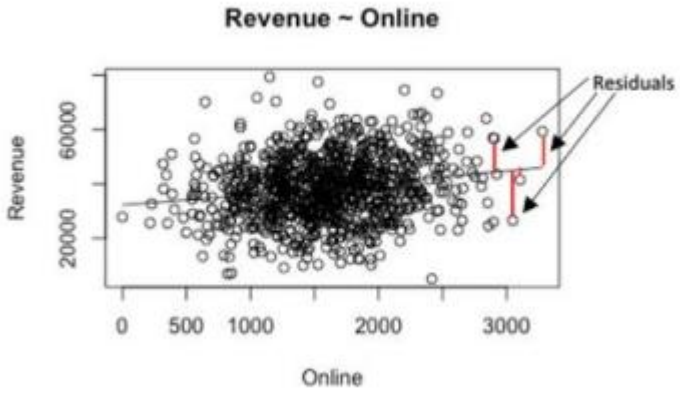

Figure 1: Linear regression

In hypothetically, it is interpreted that the amount of online investment should be correlated with the amount of store revenue. To do this, we will test the data using t test (for unknown data) or z-test (for known data) and will find the co-efficient value of x. If the co-efficient value is zero then it will be null hypothesis, alternatively it will be alternative hypothesis which will help to continue the business (Scott 2021).

To find the more precise concept, the above data can be calculated and tested with the p-value or the probability of occurrence of the event as well as it can be hypothesised that if p-value is less than 0.05 (the threshold point), the coefficient will not be equals to zero which proofs the straight relationship between the amount of the online investment and the amount of the revenue (Scott 2021). In this way, the linear regression technique is used in the business opportunity without human-driven approach and bringing the targeted revenue based on the predictor or independent factors.

"Presenting correlation as well as binding strong relationships with the datasets are great impacts of linear regression analysis which are noticeable in the business opportunity. Apart from this, this solution helps business to predict the response feature and to take the decision accurately based the trends and patterns which means a lot for the stakeholders (ResearchOptimus 2023).

Additionally, if we imagine, we can find that the above calculation is difficult to complete by human-driven approach where supervised machine learning is used to complete this calculation insightfully by applying its test and equation which is a blessing for the business world" (As shown in appendix A).

Evaluating the supervised machine learning technique:

As regression analysis has the statistical approach specifically in the simple linear regression model, it provides business the powerful calculation based on the target and predictor features (smallbusiness 2023). With this simple linear regression model, any business can make their understanding clear because of its linearity, patterns, and trends as well as its forecasting behaviour for upcoming months and years which helps to optimize the business curriculums and decisions.

The linear regression algorithms brought its success in the business area using its powerful statistical approach (smallbusiness 2023). In this approach, it has a calculation method to calculate the value of the dependent variable based on the predictor or independent factor (Investopedia 2022). From the above grocery store scenario, it is observed that targeted variable `revenue` is easily computable and predictable based the independent variable applying some test and mathematical equation which will be approachable to all stakeholders in their business. Additionally, it is very simple linear regression model that generates with only one dependent and one independent variable and helps multiple linear regression model by adding some required independent variables in the business whereas multiple linear regression has some complexities based on its patterns and trends such as multicollinearity that forms with one or more independent variables and made high correlation with each other (HAYES 2023).

“This technique has a large impact in business goals and outcomes. In accordance with business goals, this algorithm is not only predicting the store revenue but also anticipating the risk and opportunities of the business, managing business problems appropriately by using different calculations based on the features as well as creating business amenities by improving the decision-making (ResearchOptimus 2023)” (As shown in appendix B).

Ethical concerns and consideration:

As everything has some pros and cons, this technique has a few cons regarding its trends and patterns. During the hypothetical operation, sometimes this model shows data inconsistency in its features which resulting the underfitting (iq 2023). To mitigate this concern, it is considered to apply the feature selection method where specific feature will demonstrate to lead the function accurately (IBM n.d).

“Apart from this, Khushbu and Suniti (2018) state that the evaluation of this solution is also affected by the co-efficient of determination’s (R-square) value as it sometimes represents the low values. In this case, the amount of revenue will not be as an expected value for the grocery store or any business because of the weaker relationships between both dependent and independent variables. To change this circumstance of the R-square value, it is considered to add one or more compatible independent factors so that the R-square value extends adequately (Scott 2021)” (As shown in appendix C).

2nd Business problem:

Nowadays, in financial sector specially in banking area fraud transaction from credit card is one of the big issues. In this scenario, Hewapathirana (2022) describes that when a person steals or uses someone’s credit cards without informing them, it would refer as an illicit or fraud transaction. In this

case, financial service such as banks face many challenges and issues like losing customer's trust and satisfaction as well as organisation's revenue.

Technique: Unsupervised machine learning

In industries, when suspicious transaction occurs during the normal transactions, industries such as financial industries apply some algorithms or techniques to mitigate these problems. To overcome those challenges and detect fraudulent transactions, isolation forest algorithm is used in unsupervised machine learning problem to find the anomalies from the normal transactions (Waspada et al. 2020). During this operation to get the business opportunity, some features need to be collected such as bank account details, and credit card history where transaction type, transaction time, and location of the transactions will be included. To work with the isolation forest model, we use these features in the required datasets, and it runs two steps such as a random sub-samples data (training dataset) and then it divided as a binary tree.

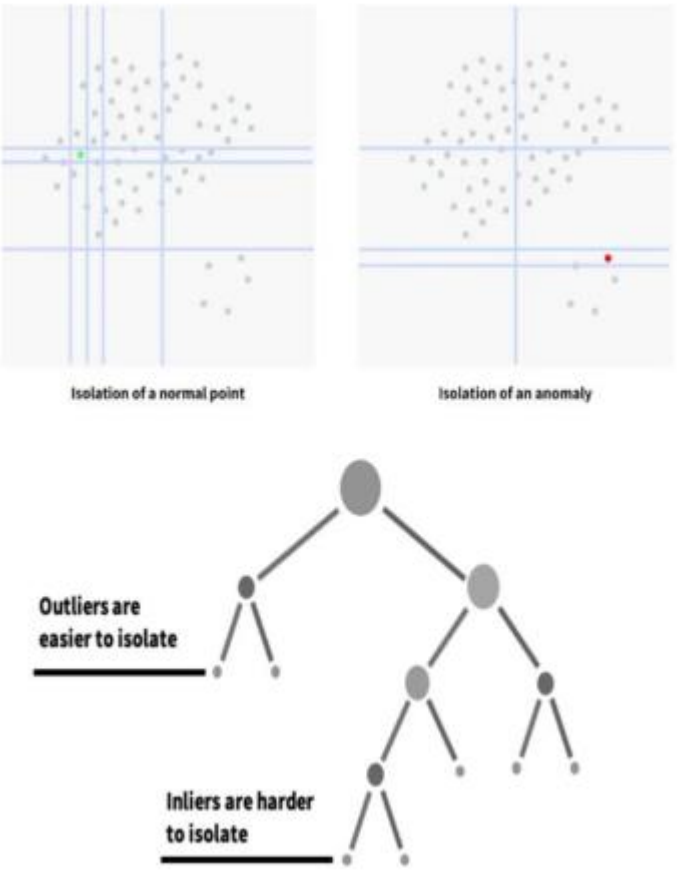

Figure 2: Anomalies detection using isolation forest model.

On that binary tree, it makes some branches where threshold values (minimum or maximum values of the selected feature) are indicated. When the value of the data is less than the selected threshold, it goes to the left or right branch of this tree. In this way, previous steps keep repeating their process in the isolation forest model until each data point is isolated to the last branch or the maximum depth which forms the training stage ((Waspada et al. 2020). Here, a data point is counted based on their scores while continuing the scores with the data point through all the trees that known as

evaluation stage (Waspada et al. 2020). In the next step, the anomaly score data point compared with the maximum depth of the data point (normal data point). In this way, anomalies or outliers or unknown behaviours are detected using the isolation forest algorithm in the business area.

"In this case, it is observed that isolation forest model is one of the unsupervised machine learning algorithms which finds the outliers rapidly and efficiently, allowing decision-makers to make data-driven decisions in real-time that might not be possible for human (Kumar 2021). Therefore, it is considered that if industries use the unsupervised machine learning model to work with unlabelled larger datasets, they will earn both trust and profit from their consumers in every year" (As shown in appendix D).

Evaluating the unsupervised machine learning technique:

As (Waspada et al. 2020) state that isolation forest model makes the sub-samples data with the training dataset, which is rare with other solutions, I agree that this technique is appropriate to work with the unlabelled larger dataset. Apart from this, McDonald (2022) suggests that it reduced the operational times and can work with larger dataset swiftly that brings the satisfaction for the users compare to the manual or human-driven work that is time-consuming and slow in earning the trust of customers and solving fraudulent issues.

Since the credit card transactions are unknown or the data are unlabelled as well as the amount of fraud data is included very imbalanced way where the new pattern of fraud data is disclosed (Waspada et al. 2020), iForest algorithm is one of the algorithms that can work with the current transactions of unlabelled large datasets and analyse the credit card fraud efficiently. In this perspective, GOMES et al. (2022) state that supervised and semi-supervise machine learning models can work only labelled and combine with label and unlabelled datasets respectively. Apart from this, Waspada et al. (2020) represent that its accuracy is also bringing the success for the industrial area specially for the financial industries which can be imagined by the collected 'validation results where precision is 0.809917, recall is 0.710145 and f1-score is 0.756757 as well as the test data results where precision is 0.807143, recall is 0.763514 and f1-score is 0.784722'. Moreover, Waspada et al. (2020) explain that isolation forest model is the most efficient technique to detect the anomalies specially for credit card fraud analysing rather than two unsupervised algorithms LOF (Local Outlier factor) and OCSVM (One-class SVM).

"It is also found that the isolation forest model obtained the best results with accuracy while it is performing the splitting and validating with the required data (Waspada et al. 2020). Kumar (2020) explains that as this technique is performing high with its results, companies or organisations are trying to apply this technique to extend their revenue by detecting the anomalies from the fraudulent activities. In this case, industries are focused on their customer satisfaction to retain their customer using the machine learning techniques where isolation forest model or iForest algorithm performed a lot to the industrial side as it has high speed, excellent scalability and less memory requirements which fulfilled the customer requirements in shortly (CHABCHOUB et al. 2022)" (As shown in appendix E).

Ethical concerns and consideration:

In this planet everything is not perfect. Though isolation forest model performs very well and has many pros, it has some cons as well where bias and inconsistency in reliability of the anomaly score are noticeable (CHABCHOUB et al. 2022).

“Ethically, it needs to consider about the i Forest (Isolation Forest Model) algorithm’s success based on its evaluation. In this perspective, it is necessary to focus on these limitations so that the unsupervised machine learning technique can success adequately. To overcome the anomaly scoring inconsistency, it is suggested to use EIF or Extension Isolation Forest which splits the data of the node randomly and fix the consistency of the scores (CHABCHOUB et al. 2022). Apart from this, we need to consider mitigating the biases using some strategies such as observing machine learning process regularly, operating and performing the model accurately as well as adding the users’ feedback mechanisms while delivering solution for the users (Vashisht 2021). After considering these mitigations, it needs to think about the cost of the mechanisms of software of the unsupervised machine learning specifically from all the angles and the estimation of the ROI (Return on Investment) so that the revenue comes adequately in accordance with the investment in every year (Linna 2020)” (As shown in appendix F).

3rd Business problem:

A small number of thousands of languages are spoken across the universe and every data is not available for every language and tongue which is identifying as a challenge for most of the organizations (altexsoft 18 March 2022). To label these large amounts of audio data manually, it must think of a very expensive and timeconsuming task which is why high-quality systems, but low-cost models need to be trained with large amounts of transcribed speech audio to make the business worthful and successful.

Technique: Semi supervised machine learning

“Nowadays, semi-supervised machine learning (SSML) i.e; combine with supervised and unsupervised machine learning is playing a vital role to overcome the language challenges and to solve the business problems effectively (VALE et al. 2021). SSML is applying self-training method to provide effective performance on those challenges. According to the working principle of self-training method, it works on small part of the data instead of adding tags to entire dataset and join with a train model, which then is applied to the lots of unlabelled data (altexsoft 18 March 2022)” (As shown in appendix G).

“During this self-training method, pseudo-labelling process is applied with the small amount of labelled data while training the model. In this stage, prediction level is high. When any pseudo level exceeds this confidence level, it will be added into the dataset and formed a new one and joined the input to train the improved model. In this way, this method keeps repeating until finding the best performance of the model.

At present, Facebook (on Meta technology) is using semi-supervised learning technique to its speech-recognition models and developed them. Here, Meta (formerly the Facebook company) is taking the base model that is trained with 100 hours of human-annotated audio data. With these

100 hours of audio data, 500 hours of unlabelled speech data are added where the self -training method is applied and increased the performance of the models (altexsoft 18 March 2022).

As the self-training method of the semi-supervised learning works with small amount of labelled data while adding with a large amount of unlabelled data, it reduces the cost of manual annotation and the data preparation time. Besides this, IBM (n.d.) indicates that this speech recognition software has "included a spelling dictionary of 100,000 words" and fast decoding algorithms which will help business to acquire it success where speech technology will be available for every customer.

As the manual labelling is costly and time consuming, semi-supervised machine learning method is helpful to suppress the problems. In a nutshell, semi-supervised learning (SSL) is a machine learning technique that uses a small portion of labelled data and lots of unlabelled data to train the predictive model where small portion of labelled data is saving times and budgets to make the business successful (altexsoft 18 March 2022). Therefore, it is observed that machine learning approach should be used instead of human-driven approach" (As shown in appendix H).

"Eventually, organisations like Meta are getting large advantages to reduce the word error rate (WER) which is decreased by 33.9% than the early days and significantly appreciable for upcoming business decisions (altexsoft 18 March 2022)" (As shown in appendix I).

Evaluating the semi-supervised machine learning technique:

Kahn et al. (2020) indicate that filtering mechanisms improves the model's accuracy by applying self-training technique which is mainly focused on WER (Word Error Rate). As the word error rate is reducing day by day, self-training method is also performing excellent in business cases in the tech world. Apart from this, only semisupervised learning method can work with the small part of the label data instead of working with the large amount of label data and iterate the pseudo labels model several times until it gets the appropriate accuracy which is unpredictable for other models like supervised or unsupervised model (altexsoft 18 March 2022). Kahn et al. (2020) explain that when self-training process iterate with its pseudolabels data, the performance of the model is increased. In hypothetically, model's performance is measuring the business success.

"In regard to business goal and outcome, some organisations are getting benefits and making some assumptions for their future business when several terms like utterance, speech, volume demonstrate accurately, and clear background noise based on WER (Word Error Rate) which leads to better customer retention as all terms are accepted to the customers (IBM n.d). Apart from this, when WER (Word Error Rate) is reduced with the model's performance, organisations also earned their revenue by completing their all tasks in real time, as employees spent less time to manage all documents using this technology (ONTASK 20 March 2017). Overall, it makes business faster and improves the entire organisations efficiency" (As shown in appendix J).

Ethical concerns and consideration:

As SSML technique has some benefits in the business area, it has limitation as well. Since it is operated with the large amount of data, it needs more operating power which makes it inapproachable for some businesses (Newton 2022).

Conclusion:

The machine learning models are profitable for most of the use cases in business. It is considered to use these methods cautiously so that organisations can make their decisions effectively for their future business success. In this article, it is discussed elaborately using three business problems such as predicting revenue, fraud detection and language recognition by applying three techniques linear regression model, isolation forest model and self-training method respectively as well as encouraged the stakeholders to make their future business decisions by following these above techniques.

References:

altexsoft (18 March 2022) 'Semi-Supervised Learning, Explained with Examples', altexsoft, accessed 16 March 2023. https://www.altexsoft.com/blog/semi-supervisedlearning/

CHABCHOUB Y, TOGBE MU, BOLY A and CHIKY R (2022), 'An in-depth study and improvement of Isolation Forest', IEEE Access, 1: 1-21.

GOMES HM, GRZENDA M, MELLO R, READ J, NGUYEN MHL and BIFET A (2022), 'A Survey on Semi-Supervised Learning for Delayed Partially Labelled Data Streams', Article in ACM Computing Surveys, 1(1): 1-2, doi: 10.1145/3523055
file:///C:/Users/Akter/Dropbox/PC/Downloads/CSUR_2022_DS_SSL_Link_to_full_paper_DOI.pdf

HAYES A (2023) Multicollinearity: Meaning, Examples, and FAQs, Investopedia website, accessed 18 March 2023. https://www.investopedia.com/terms/m/multicollinearity.asp

Hewapathirana I (2022) 'Utilizing Prediction Intervals for Unsupervised Detection of Fraudulent Transactions: A Case Study', Journal of Engineering and Technology, 11(2): 1-11, doi: 10.51983/ajeat-2022.11.2.3348.

IBM (n.d) What is underfitting? , ibm website, accessed 19 March 2023. https://www.ibm.com/topics/underfitting

IBM (n.d) What is speech recognition?, ibm, accessed 19 March 2023. https://www.ibm.com/au-en/topics/speech-recognition

iq (2023) Advantages and Disadvantages of Linear Regression, iq website, accessed 19 March 2023. https://iq.opengenus.org/advantages-and-disadvantages-of-linear-regression/

Investopedia (2022) Regression Basics for Business Analysis, Investopedia website, accessed 17 March 2023. https://www.investopedia.com/articles/financialtheory/09/regression-analysis-basicsbusiness.asp#:~:text=Simple%20linear%20regression%20is%20commonly,could%20affect%20sales%2C%20for%20example

Khushbu K and Suniti Y (2018) 'Linear Regression Analysis Study', Journal of the Practice of Cardiovascular Sciences, 4(1): 1-5, doi: 10.4103/jpcs.jpcs_8_18.

Kumar S (2021) 5 'Anomaly Detection Algorithms every Data Scientist should know', towardsdatascience website, accessed 19 March 2023. https://towardsdatascience.com/5-anomaly-detection-algorithms-every-data-scientist-should-know-b36c3605ea16#:~:text=Anomaly%20detection%20algorithms%20are%20very,anomalies%20from%20the%20training%20sample

Kumar D (2020) 'Fraud Analytics using Extended Isolation Forest Algorithm' linkedin website, accesses 19 March 2023. https://www.linkedin.com/pulse/fraud-analytics-using-extended-isolation-forest-algorithm-kumar/

Linna E (2020) 'Return on Investment for Machine Learning', towardsdatascience website, accessed 17 March 2023. https://towardsdatascience.com/return-on-investment-for-machine-learning-1a0c431509e

McDonald A (2022) 'Isolation Forest – Auto Anomaly Detection with Python', towardsdatascience website, accessed 20 March 2023. https://towardsdatascience.com/isolation-forest-auto-anomaly-detection-with-python-e7a8559d4562#:~:text=Advantages%20of%20Isolation%20Forest&text=Reduced%20computational%20times%20as%20anomalies,when%20irrelevant%20features%20 are%20included

Newton E (2022) 'You Need to Know the Pros and Cons of Self-supervised Learning', itchronicales website, accessed 24 March 2023. https://itchronicles.com/artificial-intelligence/you-need-to-know-the-pros-and-cons-of-self-supervised-learning/#:~:text=One%20of%20the%20most%20significant,it%20inaccessible%20for%20some%20teams

ONTASK (20 March 2017) 'The Cost of Paper Processes in the Workplace', ontask, accessed 23 March 2023. https://www.ontask.io/resources/blog/the-cost-of-paper-processes-in-the-workplace/

ResearchOptimus (2023) Correlation and Regression Analysis Aiding Business Decision Making, ResearchOptimus website, accessed 16 March 2023. https://www.researchoptimus.com/article/what-is-correlation.php

Scott J (2021) Simple Linear Regression for Business, medium website, accessed 15 March 2023. https://medium.com/@jansco/simple-linear-regression-for-business-346566d011cb

smallbusiness (2023) The Advantage of Regression Analysis & Forecasting, smallbusiness website, accessed 16 March 2023. https://smallbusiness.chron.com/advantages-regression-analysis-forecasting-61800.html

VALE KMO, GORGONIO AC, GORGONIO FDLE, and CANUTO AMDP (2021), 'An Efficient Approach to Select Instances in Self-Training and Co-Training Semi-supervised Methods', IEEE Access, 1: 1-23, 10.1109/ACCESS.2017.doi.
file:///C:/Users/Akter/Dropbox/PC/Downloads/An_Efficient_Approach_to_Select_Inst ances_in_Self-.pdf

Vashisht R (2021) 'How to reduce machine learning bias', medium website, accessed 18 March 2023. https://medium.com/atoti/how-to-reduce-machine-learning-bias-eb24923dd18e

Waspada I, Bahtiar N, Wirawan PW and Awan BDA (2020), 'Performance Analysis of Isolation Forest Algorithm in Fraud Detection of Credit Card Transactions', Jurnal Ilmu Komputer dan Informatika, 6(2): 1-11