**ICT Project Assignment**

**Task 1**

**Introduction**

The main area of study proposed is Machine Learning. Machine learning is a branch of artificial intelligence which is centred on the concept that system can learn from data, gain the ability to identify patterns, thus leading to decisions with minimal human intervention. The iterative position in machine learning is crucial as models are subjected to new data, they adapt in an independent matter (Machine Learning: What it is and why it matters, 2021).

In this research, the research will focus on supervised machine learning. Supervised machine learning is the pursuit for algorithms that use externally supplied occurrences to reason and create a general hypothesis, which then make expectations for future instances. The goal in supervised learning is to construct a short model of the distribution of group labels in predictor feature conditions. The classifier of result is then used to designate group labels in the testing occurrences where the values of the predictor features are recognized, yet the significance of the group level is unknown (Kotsiantis, 2007).

**Completed Work**

* Accounts were created within the first week of September 2021 and throughout September and October the hypothesis, research questions, research methods, research pipeline, and finalization of the report were concluded.
* The literature review assets and algorithms were viewed within the first two weeks of November. Evaluation metrics and the finalization of the report were concluded by the end of November.
* In the literature review, the research carried out was to create a machine learning environment in which car dealership prices and sales in Malta are analysed and learned. Ultimately this leads to the machine predicting the future sales of car dealership cars, with the use of supervised regression learning. This will aid risk management and provide advantages and disadvantages of important business decisions, as well as understanding the organization’s current situation and future situations to make the correct business decisions.
* The prototype was then attempted from the beginning of December, research was made on tutorials and relevant datasets up. The prototype applied for the intent of this research is centred on exhibiting that a machine learning technique can eventually forecast upcoming sales figures.
* The prototype has ample functionalities, loading, cleaning, and handling data from the dataset provided, assign the model’s dataset in the parameters of the model which reads, understands, and learns the dataset to predict the selected data’s price. The standalone application is desktop based, meaning that there is no need for internet. The only use of internet is to install libraries and packages to be pre-installed and loaded.
* In total the prototype took around 2 full months to finish and apply to the relevant study until the end of January 2022.
* Finally, the questionnaire has been made via SurveyMonkey, to be given to car dealers in order to get a better understanding of how the prototype is seen in their experienced opinion.

**Work to be Done**

The rest of the work to be done are:

* Collection of questionnaires from the car dealers which takes approximately 1-2 weeks.
* Evaluations of the prototype, the implementation, the steps needed for the prototype to function, pilot testing, the errors and their description, and the ethical considerations along with these.
* Evaluations of the dataset, what can be derived from the collected questionnaires and how will this data affect the study.
* The hypothesis will then be tested via the relevant formulas.
* The contributions made by this study will involve additional insight to the effects of implementations and understanding of future new car markets from auto dealers in Malta. This study will help the understanding of the upcoming new car sale sales and the decision changes auto dealers are prone to with the use of machine learning prediction.
* The report will be formulated, followed by the conclusion and recommendations, introduction, and abstract which will all outline the purpose of the study.

**Conclusion**

In this Interim Project Report the main area of study has been highlighted within the introduction. The research focus and the relevance to supervised machine learning has been outlined. Following these, the work done so far has been outlined and described thoroughly, while the work to be done has been highlighted and summarized.

**Task 2**

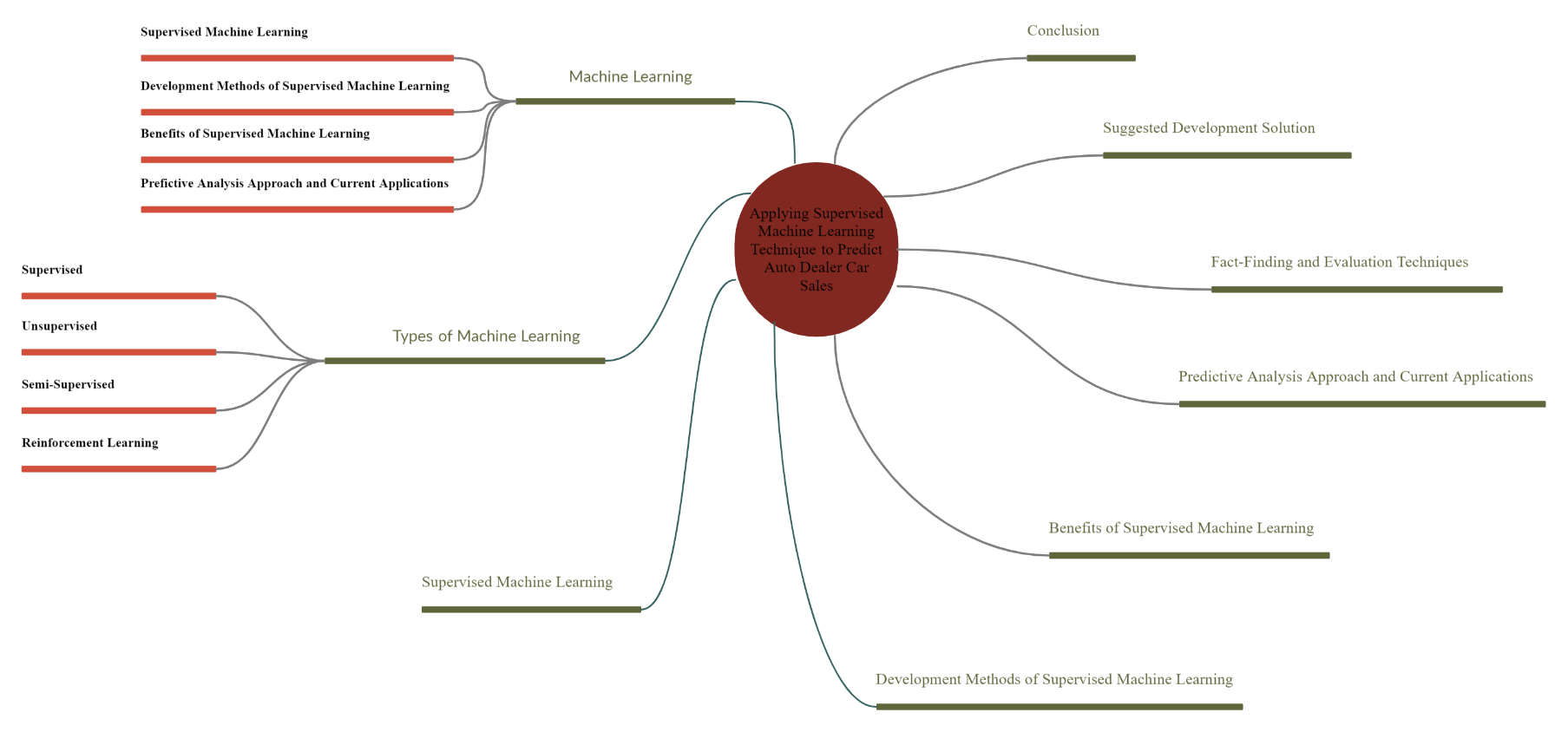


Figure 1 – Concept Map

The concept map only outlines the major parts of the study. These are the main focuses of what the study shall be based on without giving too much information. The limitations of this concept map are that this does not reflect the full pattern of thought throughout this structure process. Certain points can be difficult to understand from the concept map itself and are easier to grasp from the actual description rather than just the title. This concept map outlines the ‘bigger picture’ of the study and visualises relationships. Changes were made to this concept map all throughout the transition of the dissertation as certain concepts were understood and irrelevant data was omitted.

**Task 3**

1. Machine Learning Model for Sales Forecasting by Using XGBoost - The purpose of this case study is to develop a design, implementation, and evaluation of ForeXGBoost. ForeXGBoost is a program which takes benefit of carefully designed data filling algorithms missing values to enhance data quality. This case study is being done for various industries where prediction models are built to forecast sales in order to aid planning. At this stage in the research, the ForeXGBoost will be generally defined as a supervised machine learning method (dairu and Shilong, 2021).
2. SemiBoost: Boosting for Semi-Supervised Learning - The purpose of this case study is to improve the classification accuracy of any given supervised learning algorithm by using the available unlabelled examples. This SemiBoost algorithm packages around the fundamental supervised algorithm and improves its performance using unlabelled data for organisations which make use of predictive machine learning programs in order to predict where the business will flow. At this stage in the research, the SemiBoost algorithm will be generally defined as a semi-supervised machine learning algorithm (Mallapragada et al., 2009).
3. Used Car Pricing and Beyond: A Survival Analysis Framework - The purpose of this case study is to understand the lifetime of used car listings based on a survival analysis approach. This can then be used for auto dealers or relevant organizations on websites such as Autotrader. At this stage in the research, the used car pricing predictor will be generally defined as supervised machine learning algorithm. (Demiriz, 2018)

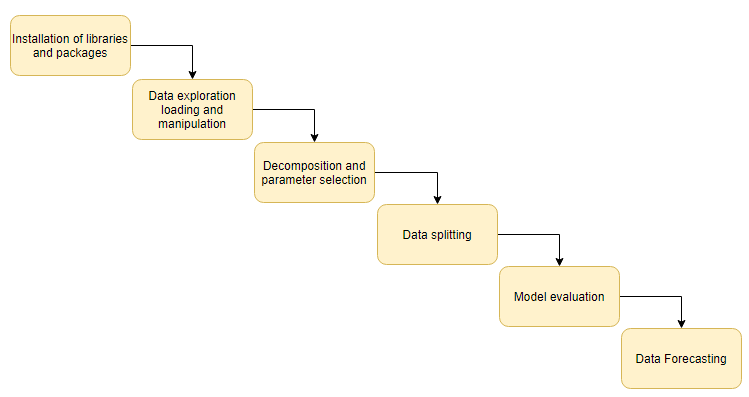
**Task 4**

1. The research design used in this study is quantitative research. As opposed to qualitative research, quantitative research aims to measure a percentage or figure. Moreover, quantitative research constructs a statistical model, classifying and counting features, resulting in a more refined explanation of the observed (McCusker and Gunaydin).

Once data is gathered, it is to be processed before analysis. For example, survey and test data may need to be transformed from words to numbers. Then, statistical analysis is used to answer your research questions. You can also evaluate the consistency and legitimacy of the data collection methods to indicate how consistently and accurately the methods actually measured what their initial intention was for. Quantitative research templates are factual, detailed, and investigational. The results attained from this research method are rational, statistical, and neutral.

1. Quantitative research is regularly used to normalize data collection and generalize conclusions. Organized tools such as surveys, polls, or questionnaires are used to collect quantitative data. Using such structured methods facilitates the collection in depth and actionable data from the survey respondents. Hypothesis testing, using validated and determined hypothesis testing procedures means that research variables, predictions, data collection and testing methods are to be precisely considered and reported before concluding. This research method also collects reliable and accurate data, since data collected in numbers is accurate, a precise representation of the conducted research is offered without discrepancies.

Figure 2: Prototype implementation pipelines



As shown in figure 2, execution of the prototype was split into six stages. This prototype included different experiments and included multiple different scripts. For the dataset ten columns were removed from the dataset in order to keep relevant data for the model prediction. The six stages of implementation were as follows:

**Installation and loading of libraries and packages**

The necessary packages have been installed to run the packages necessary for the program to be able to generate the model. Packages such as numpy, matplotlib, statsmodel.api, etc. These were used for different sections of the prototype such as graphs, data explanation, and specific processes. These packages are necessary for the program to function.

**Data exploration, loading and manipulation**

The dataset was loaded onto the Jupyter notebook and has been enhanced by removing unnecessary columns for the model to understand and make use of the relevant data as well as removing all the null values. The null values have been removed in order to avoid inaccurate data and errors this will cause in the prototype in future testing.

**Decomposition and Exploratory Data Analysis**

After the cleaning of the dataset, pandas profiling and matplotlib packages were used to generate graphs of various columns and observe their relationships with the target column. This helps in understanding of readings of the data provided while also grasping a better concept of the dataset via exploratory data analysis with only some lines of code.

**SARIMAX**

Here is when Seasonal Auto-Regressive Integrated Moving Average with external factors (SARIMAX) in Process Data Quick (PDQ) models in combination with Seasonal PDQ has been used for the dataset. These focus on seasonal effects and external factors with the autoregressive and moving average component in the model. The primary use of this autoregressive model is due to its capability of predicting future behaviour based on past behaviour, which allows easy and accurate forecasting with correlation between values in the time period which precede and succeed the model.

**Model evaluation**

The model then shows multiple visual graphs in a figure, namely the sales figures, trend line, seasonal sales figures, and residual values. Following these values calculations are made by the program in order to compare the values between observed and expected values. This is shown in the standardized residuals for ‘s’, a histogram of these standardized residuals including estimated densities, sample quantities, and a correlogram. This model provided the estimated accuracy of the model currently being used for this specific dataset. The next figure visually portrays the actual data as well as the one-step ahead forecast’s estimation. Lastly, the final figure shows the actual data, followed by the forecasting of the future 2 years after the dataset has concluded.

References:

1. Sas.com. 2021. Machine Learning: What it is and why it matters. [online] Available at: <https://www.sas.com/en\_us/insights/analytics/machine-learning.html#:~:text=Machine%20learning%20is%20a%20method,decisions%20with%20minimal%20human%20intervention.> [Accessed 1 July 2021].
2. Kotsiantis, S.B., Zaharakis, I.D. and Pintelas, P.E. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, [online] 26(3), pp.159–190. Available at: <http://www.informatica.si/index.php/informatica/article/viewFile/148/140> [Accessed 24 Mar. 2019].
3. dairu, X. and Shilong, Z. (2021). *Machine Learning Model for Sales Forecasting by Using XGBoost*. [online] IEEE Xplore. Available at: <https://ieeexplore.ieee.org/abstract/document/9342304>
4. Mallapragada, P.K., Jin, R., Jain, A.K. and Liu, Y. (2009). SemiBoost: Boosting for Semi-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11), pp.2000–2014.
5. Demiriz, A. (2018). *Used Car Pricing and Beyond: A Survival Analysis Framework*. [online] IEEE Xplore. Available at: https://ieeexplore.ieee.org/document/8665680 [Accessed 30 Sep. 2021].