**Chapter 3**

This chapter will discuss the research methodologies validity and reliability with relation to the topic chosen. Additionally, this chapter will include the research strategy, data collection methods and tools, prototype, pilot testing, errors, ethical consideration, and conclusion.

**3.1 Research Strategy**

The research held with respect to this study was based on deductive reasoning. This strategy is evident since the study is initialized with a hypothesis, in this case “Applying Supervised Machine Learning Technique to Predict Auto Dealer Car Sales” where a logical conclusion is reached via studying and examining possibilities. The strategy proposed has unfolded into new research with previously existing context.

Deductive reasoning is based on deciding on a premises that is generally assumed to be true. As stated in Liew et al. (2018), the process of deductive reasoning is one that generates conclusions which either do or do not necessarily come from a given premises. Deductively valid arguments induce conclusions that ought to be true, provided that the premises are true. In deductive reasoning multiple means such as accepted truths, theories and laws are used to prove the accuracy of the conclusion.

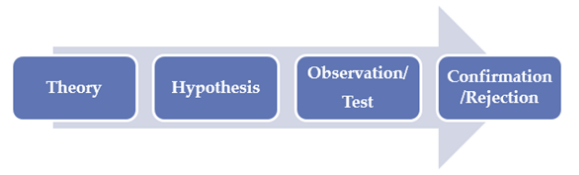


Figure 3.1 – Deductive reasoning

**3.1.1 Research Method**

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.

**3.1.2 Research Method Justification**

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.

**3.2 Data Collection Methods and Tools**

Data statistics and analysis were performed on the data gathered in this research. Methods in collection of data consists of an online survey and experiment. An online survey was required to collect the necessary information to perform the cross tabulation, descriptive and inferential statistics, all based on the research question “Does applying supervised machine learning technique increase auto dealer car sales?” The target participants were vehicle auto dealers. In this case with a population size of 116 auto dealers (Car Dealers in Malta & Gozo, 2021), a 10% margin of error, and a 95% confidence, the sample size is of 53 respondents. These calculations have been made using Raosoft, shown in figure 3.2 below.

An experiment in which a solution was implemented was used for the second data collection method, and by using the evaluation metrics mentioned the results were correctly tested. Data collection has been included in this experiment which pertained the second research question: “Can machine learning using a supervised learning method accurately predict auto dealer sales?”

Graphical user interface, text, application, Word

Description automatically generated

Figure 3.2 – Raosoft Sample Size calculator

**3.2.1 Advantages and Limitations of the chosen Data Collection Tools**

Benefits and limitations are brought with the data collection method chosen. Online surveys proved to have multiple benefits during this research. According to Wright (2006), costs in hardware and software have significantly decreased in comparison to older times thus leading to less costs than traditional methods. Organisations now aggressively promote their presence, yet these not only offer information to consumers, these present opportunities for researchers to access varieties of these populations. Such online presence provides much easier communication for these surveys, where instead of physically presenting oneself for each participant, they can be all targeting and communicated online. Simplicity is also a factor since this is much easier for participants to participate and for the researcher to tabulate and visualise.

Limitations in online surveys were evident as well, since the focus of the survey increased on predictive machine learning this might progressively lose the interest of the participants and they can either fill in these randomly or refrain from completing the survey. Another limitation is these results may have been subject to bias which would lead to random or potentially fake responses. As a result, the outcomes of the survey may not have been as accurate as expected.

Experimental data collected also presented benefits as well as limitations. Multiple free online tools have provided access to different datasets used to fulfil different operations. However, it is notable that each dataset cannot contribute directly to the operation taking place in this research and these datasets require transformations to make them useful for the purpose of this research. This results in a very time-consuming operation, yet when these are cleaned and updated, not much time was required to fulfil the purpose of the experiment.

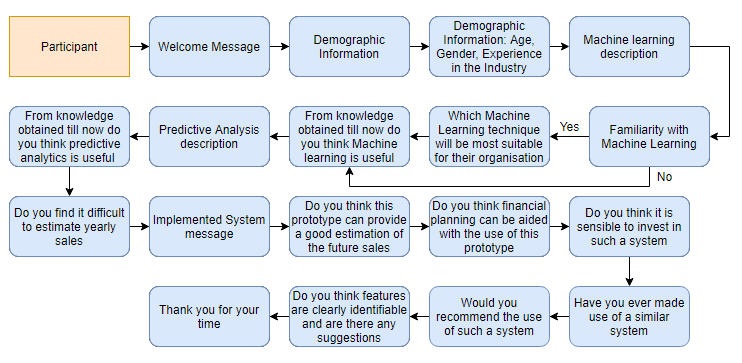
**3.2.2 Data Collection Methods and Tools Structure and Information**

As can be seen in appendix 3.3 the online survey has four sections. The first section included the title, information about the researcher, the field of study and commitments regarding the material provided by the participants. The second section, demographic information is asked from the participants, including age, gender and experience in the industry. Additionally, filtering by selected demographic information this has served as a great base for statistics of results.

Five questions were present in the next section, starting off with a small description of Machine learning; the familiarity of the participant with machine learning, consisting of a follow-up question of which Machine Learning technique would be most suitable for the organisation, if answered yes; and from the knowledge obtained so far if Machine Learning is useful in the automotive industry. A small description of predictive analytics is provided prior the fourth question, since this comprises of the benefits of predictive analytics; the fifth question is regarding the difficulty of estimating number of sales per year.

The fourth section consists of six questions about the experiment presented. This section asks the participant whether the experiment provides good sight of the future in predictions, financial planning aid and if such systems would be used for their organisation.

Figure 3.3 - Online survey schematic diagram



**3.3 Prototype**

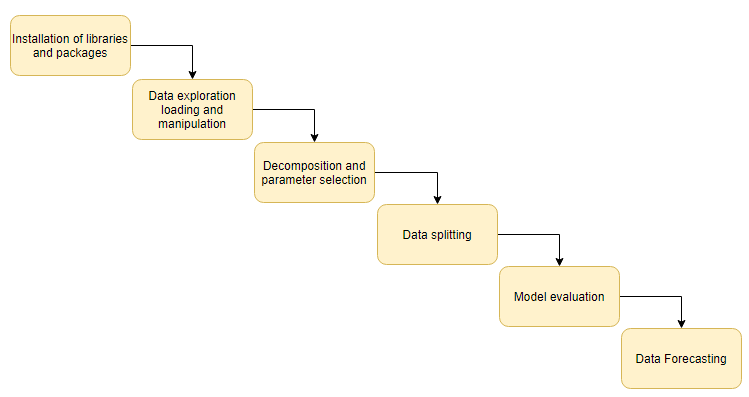
The prototype applied for the intent of this research is centred on exhibiting that a machine learning technique can eventually forecast upcoming sales figures. The prototype has ample functionalities, loading, cleaning, and handling data from the dataset provided, assign the model’s dataset in the parameters 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.

**3.3.1 Development Tools**

Various tools were used for this application. The programming language used was Python 3.10 alongside the Jupyter notebook. Moreover, several third-party programs, libraries, and packages such as: numpy (version 1.21.3) used for mathematical calculations, matplotlib (version 3.4.3) used for graph plotting, pandas (version 1.3.4) assisted in data loading and conversion of data, statsmodels.api (version 0.13.0) for statistical data exploration and Postman, used for API testing. It is an HTTP client that examinates HTTP requests, using a graphic user interface, through which we acquire several types of reactions that need to be validated after. As for datasets, Microsoft Excel was used to edit, read, and export comma-separated value (csv) files. No hardware was used for the implementation of this prototype.

**3.3.2 Implementation**

Figure 3.4: Prototype implementation pipelines



As shown in figure 3.4, 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:

**Stage 1: 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.

**Stage 2: 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.

**Stage 3: 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.

**Stage 4: Catboost Regression**

Here is when catboost regression model has been used for the dataset. Catboost focuses on gradient boosting in decision trees, which skips the step of label encoding categorial features since catboost can work on categorial features directly. Catboost was installed then a list of categorial features were passed over to the model and trained on the dataset.

**Stage 5: Model evaluation**

The model accuracy was then checked via the Root Mean Square Error (RMSE) model. This model provided the estimated accuracy of the model currently being used for this specific dataset. Depending on the parameters used the RMSE value was varied according to the features of the dataset put into the catboost model. The dataset used and the model predictions accuracy were based on the RMSE value. The model was then saved into a pickle file.

**Stage 6: Data Forecasting**

A webpage has been created for the frontend side of the prototype in order to select details of the forecast, depending on the selected data, the forecast will change accordingly as the model will re-evaluate the data based on selection. The forecast is then made on the given webpage and the model will forecast the sales for a specific car according to the details provided.

**3.4 Pilot Testing**

Prior to the final version release of the prototype developed, pilot test testing was performed in case of any syntax errors, logical errors, or any improvements to be made.

Table 3.5: Data loading testing

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Action | Inputs/Function used | Expected | Actual | Pass |
| The data was loaded from the correct path in the excel worksheet. | read\_csv(@correctpath) | From the excel worksheet, data is read, then placed into a data frame. | Error indicated that an incorrect function was used. | No |
| The data was loaded from the correct path in the excel worksheet. | read\_excel(@correctpath) |  |  | Yes |
|  |  |  |  | Yes |

**3.5 Errors**

**3.6 Ethical Considerations**

The researcher has created certainty that only himself, the supervisors and examiners have access to all the information collected by not providing admission of the study findings to anyone but the permitted personnel. The design of the survey provides an anonymity throughout both participation as well as data the input of data. Names, surnames, addresses, emails, and personal information have not been taken in this survey. Participants were not involved in any physical harm as the information was gathered online and at the convenience of the participant. Regarding moral harm, this was greatly avoided due to a short, accurate video regarding key technical words was provided to better understand the subject. Moreover, questions in the survey were straight to the point for participants to easily understand what is being asked from them. Ultimately, business harm was omitted as no specific business name or information was used throughout this research.

**3.7 Conclusion**

This chapter discussed the research methodologies validity and reliability with relation to the topic chosen. Additionally, this chapter included the research strategy, data collection methods and tools, prototype, pilot testing, errors, ethical consideration, and conclusion. In the next chapter the analysis of results will be discussed in detail and will include the data analysis method, analysis and discussion of the online survey, analysis and discussion of the experiment, analysis, and discussion in relation to the literature and analysis and discussion in relation to the hypothesis and research questions.