# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Ronik Jayakumar |
| **Project Name** | Assessment 2 Experiment 1 |
| **Date** | 13/04/2024 |
| **Deliverables** | AT2 Experiment 1  Support Vector Machine |

|  |  |
| --- | --- |
| 1. **EXPERIMENT BACKGROUND** | |
| Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach. | |
| **1.a. Business Objective** | Explain clearly what the goal of this project for the business is. How will the results be used? What will be the impact of accurate or incorrect results?  We are working for a car reseller company who is planning their next marketing campaign. The goal of the described project is to build a model that can accurately predict what segment of customers are most likely to buy another car so that marketing efforts and resources can be efficiently used on the right customer category.  The results of the project will be used to identify customers who fall under a defined category based on certain business questions asked and existing data on customers. The model will be run on this data to identify customers of a certain category who are most likely to make another car purchase so that marketing efforts can be used on them.  The impact of accurate results would be correct customer segmentation and target group identification which would result in high sales turnover.  On contradiction, the impact of inaccurate results would result in wrongful customer identification which would allow marketing resources to be expelled on customers not likely to make purchase, thereby reducing profitability. |
| **1.b. Hypothesis** | Present the hypothesis you want to test, the question you want to answer or the insight you are seeking. Explain the reasons why you think it is worthwhile considering it.  The hypothesis that is being tested is to check if past trends in car maintenance can predict the correlation with a customer’s inclination to buy a second car.  The insights required relate to what are the most useful questions to put forth to customers that would help us accurately find the customer segment most inclined towards investment into a new car.  It is worthwhile considering this hypothesis and investing into this model as it would help reseller company make informed data driven decisions. A company could have hundreds of thousands of potential customers on their database. Identifying the right customer segment so that marketing efforts are rightly targeted is key.  Having an accurately working model can ensure right segments are being targeted which in turn would increase sales turnover and overall revenue. |
| **1.c. Experiment Objective** | Detail what will be the expected outcome of the experiment. If possible, estimate the goal you are expecting. List the possible scenarios resulting from this experiment.  The expected outcome of this experiment is a model that can detect, classify, and predict a certain segment of customers, based on past car related activities, on if they are likely to invest into a second car.  An estimation of the expected goal would be a model that achieves at least an 80% overall accuracy in ensuring the right customer group is identified.  The possible scenarios resulting from this experiment are as follows:   * Successful Prediction – The model identifies the right customer, and the customer makes a purchase. * Missed Opportunity – The model fails to identify a potential customer, and the customer ends up purchasing a car. * Successful Rejection – The model correctly identifies an unlikely to purchase customer. * False rejection – Model identifies a wrong customer as someone who is likely to purchase a car which results in wasted marketing efforts. |

|  |  |
| --- | --- |
| 1. **EXPERIMENT DETAILS** | |
| Elaborate on the approach taken for this experiment. List the different steps/techniques used and explain the rationale for choosing them. | |
| **2.a. Data Preparation** | Describe the steps taken for preparing the data (if any). Explain the rationale why you had to perform these steps. List also the steps you decided to not execute and the reasoning behind it. Highlight any step that may potentially be important for future experiments.  Data preparation has been a key part of the process of the overall project. The given dataset is explored into to check for null values, data types, and correlation between the columns and the target dataset.  Based on the findings the following steps have been taken:   * The column titled ‘age\_band’ has been dropped due to having 112,375 null values out of a total of 131,337 values. * The data distribution of the car\_model column is looked at next. The column shows high representation bias of the different car models that exist. Due to high underrepresentation of some models, and the next column being able to give us a good picture on types of cars, the column has been dropped to decrease model complexity. * The car\_segment column consists of 4 unique features with the ‘Other’ feature having very low representation compared to the other 3. Due to no specific details being know on what the ‘Other’ column represents, the 58 rows are dropped. * The gender column consists of half the overall values being null. The column was not dropped as it is important for our business as we are identifying potential customers where gender plays an important role. To deal with this, we have used the Multivariate Imputation by Chained Equations (MICE) technique. Using this, the missing values have been identified, encoded, and predicted into a new column named gender\_no. * Using Label Encoder, the car\_segment column has been encoded into integer values to facilitate modelling.   The steps that have not been taken are   * Standardization: This step has been skipped due to the dataset already being in deciles. We have also skipped performing standardization on the gender\_no and car\_segment columns to maintain their significance. * Class Imbalance handling: This experiment is to merely check the difference between the workings of basic classification models as compared to more deeper and comprehensive ones.   Steps that could be potentially important for future experiments:   * Collect and deploy more data on customers who made a purchase of a second car. The existing data has a high class imbalance between customers who did not make a purchase and the customers who did, corresponding to the 0 and 1 values in the target column. * More inclusion: Ensure more inclusion in data collection techniques to ensure representation of people from all aboriginal cultural fronts. |
| **2.b. Feature Engineering** | Describe the steps taken for generating features (if any). Explain the rationale why you had to perform these steps. List also the feature you decided to remove and the reasoning behind it. Highlight any feature that may potentially be important for future experiments.  Feature Engineering has been performed in the dataset.   * **Gender Column**: The column contained approximately 50% of all the values as null with the other 50% corresponding to male or female. Since a high number of rows were null, dropping them would mean getting rid of a large amount of useful data. This has been tackled using Multivariate Imputation by Chained Equations or MICE for short. The following steps were taken.   + The entire column was first encoded, followed by which the Nan class was reverted to its original form to make sure null values get predicted.   + The imputing was done into a new column gender column where all the null values were predicted.   + The values were mapped back to their base form, that is as Male and Female. * **Feature Selection**: Feature selection techniques have been performed to decide what features have the highest relation to the target column. The car\_segment and the gender\_no columns are chosen to ensure representation of people and the type of car. Along with this the following was performed:   + **Correlation Analysis:** A correlation matrix was calculated between all the columns and the target column. The correlation of each column with the target column was visualized using a heatmap.   + **ANOVA-F test**: The ANOVA test was carried out as well as a second check on the correlation analysis. Once the features with the highest influence were identified, they were isolated and a new DataFrame was created with these features along with the 2 chosen categorical features. |
| **2.c. Modelling** | Describe the model(s) trained for this experiment and why you choose them. List the hyperparameter tuned and the values tested and the rationale why you choose them. List also the models you decided to not train and the reasoning behind it. Highlight any model or hyperparameter that may potentially be important for future experiments.  The models trained for this experiment are as follows:  Baseline model: **Logistic Regression**  Logistic regression has been chosen as the baseline model in this experiment due to the following reasons:   * Easy interpretability of the influence of each feature on the target column. * It is one of the most efficient base models for handling large datasets. * Gives us a good idea on the performance of the dataset on base models before going ahead with more advanced techniques.   Main model: Support Vector Machine (SVM)  SVM was the chosen model for experiment 1. This was due to the following reasons:   * SVM has a simple way of classifying data on an n dimensional plane. Its ability to perform well in high dimensional spaces make them useful for our dataset. * The kernel trick in SVM can be used to handle nonlinear data using kernel functions. * SVM has great regularization parameters to help control the trade-off between maximizing the margin and minimizing classifying errors, which lead to better general performance.   **Hyperparameters tuned:**  Hyperparameter tuning was done for the Support Vector Classifier using GridSearchCV. The hyperparameters investigated were as follows:   * C – parameter that trades off misclassification of training examples against the simplicity of the decision surface.   + Values of C investigated: 0.1, 1, 10. * Gamma – the parameter that defines how much influence a single training parameter has.   + Values of gamma investigated – 1, 0.1, 0.01. * Kernel – parameter that defines the type of kernel to be used.   + Kernel values investigated – linear, radial bias function.   The results of the best hyperparameters for support vector classifier for our dataset were as follows:   * C = 10 * Gamma = 0.1 * Kernel = rbf   The models that we have currently chosen to not train are other tree-based models such as decision trees and random forest to first get an idea on the performance of the support vector machines. Insights gained from this model could be highly useful in ensuring other models work well.  Hyperparameter tuning for future experiments could be done with higher computational power to investigate and find better values for the svm model. Limitations in computational power and speed meant choosing a smaller array of values to explore. |

|  |  |  |  |
| --- | --- | --- | --- |
| 1. **EXPERIMENT RESULTS** | | 1. **tipl** |  |
| Analyse in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified. | |  |  |
| **3.a. Technical Performance** | Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes.  The final performance results were as follows:  Training classification report:  precision recall f1-score support  0 0.99 1.00 0.99 89427  1 0.91 0.58 0.71 2468  accuracy 0.99 91895  macro avg 0.95 0.79 0.85 91895  weighted avg 0.99 0.99 0.99 91895  Validation classification report:  precision recall f1-score support  0 0.99 1.00 0.99 19181  1 0.87 0.56 0.68 511  accuracy 0.99 19692  macro avg 0.93 0.78 0.84 19692  weighted avg 0.99 0.99 0.98 19692  Testing classification report:  precision recall f1-score support  0 0.99 1.00 0.99 19151  1 0.88 0.57 0.69 541  accuracy 0.99 19692  macro avg 0.93 0.78 0.84 19692  weighted avg 0.98 0.99 0.98 19692  As we can see above, the model performed very well on the majority class 0 in all the three models trained.  **Underperforming metrics:**  The model’s evaluation metrics significantly decreased with the minority or the 1 class in the target column.   * The precision for all the three tests were good at 0.91 for training, 0.87 for validation and 0.88 for testing which indicates that a high number of values predicted as class 1 were class 1. * The recall scores for all the three models could use significant improvement with scores of 0.58 for training, 0.56 for validation, and 0.57 for testing. This indicates the model misses a high number of instances where it needs to predict Class 1. * The low recall score brings down the overall f1-score to 0.71 for training, 0.68 for validation and 0.69 for testing. * The overall accuracy is high for all the three models at 0.99 but this is influenced majorly by class 0. |  |  |
| **3.b. Business Impact** | Interpret the results of the experiments related to the business objective set earlier. Estimate the impacts of the incorrect results for the business (some results may have more impact compared to others)  The results of the experiment based on the business objectives put forth are as follows:   * The model does a good job in predicting customers who are not a potential of sale. This indicates marketing efforts can be saved on these customers and efforts and resources will be better utilized. * The model does not do a very good job in predicting potential customers which would indicate an overall achievement of 7/10 potential customers predicted accurately.   The impact of incorrect results could indicate loss in customers due to missed marketing opportunities thereby indicating loss in revenue. |  |  |
| **3.c. Encountered Issues** | List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them. Highlight also the issues that may have to be dealt with in future experiments.  Issues faced during the experiment are as follows:   * Null values in age range   + We do not get a good idea on the age ranges of people performing repurchases due to a large portion of the dataset being null.   + The path chosen in this case was to drop the column. Future steps to solve the problem would be to ensure more effort is taken into capturing customer details.   + Details such as age could be made standard to ensure no data loss. * Gender null values   + Gender could be a point of discussion due to potential privacy issues arising.   + In the case of our dataset, we can assume that most null columns were left blank intentionally by customers due to the lack of options beyond male or female. This could prove to be difficult in classifying people into segments.   + For our experiment, the model has been trained using MICE to iterate over all the null values and replace them with male or female.   + Future steps would require understanding of how details can be segmented better, like the existence of an inclusive segment. * Repetitive columns   + We have not taken into consideration columns such as nonscheduled paid services and nonscheduled service warranty. Our business case is to identify customers of a certain car type who would make a repurchase.   + A big reason for leaving this out is the fact that a lot of nonscheduled services go undocumented. We have also left it due to it showing low correlation with the target value. So to reduce processing time for the model, we have decided to leave it out.   Future problems:   * Future problems could arise in the following ways:   + Privacy issues: Data collection on car maintenance should comply with privacy laws and regulations. Privacy, data storage, and data management could a problem.   + Underrepresentation of Aboriginal culture: The aboriginal culture is a wide array of multiple nations and clans with different cultural beliefs and aspects. Ensuring proper representation and consideration into the collection, processing, and storage of data needs to align with the custodians of the land. Ensuring proper research and groundwork is done before going ahead with the project would be avid. |  |  |

|  |  |
| --- | --- |
| 1. **FUTURE EXPERIMENT** | |
| Reflect on the experiment and highlight the key information/insights you gained from it that are valuable for the overall project objectives from a technical and business perspective. | |
| **4.a. Key Learning** | Reflect on the outcome of the experiment and list the new insights you gained from it. Provide rationale for pursuing more experimentation with the current approach or call out if you think it is a dead end.  The outcome of the project has proved to be simple yet effective. We have learnt the following points:   * Effect of Deeper model: The main goal of experiment 1 was to see the difference in working between basic models and advanced models. The change in accuracies and pattern recognition is visibly apparent between the chosen baseline Logistic Regression model and the more advanced Support Vector Machine model. **This would rank very highly.** * Effect of class Imbalance: The effect of a skewed dataset is represented in this experiment. We can see the difference between the prediction numbers of the majority class 0 and the minority class 1 target variable. **This again would rank highly.** * Hyperparameter tuning: We see that hyperparameter tuning is also a major contributor towards ensuring we use our machine learning model to the best of its ability. Tuning hyperparameters has given us the most relevant attributes to apply into our model to ensure maximum extraction. **This would be of a medium rank as the effectiveness** of this step depends on the deployment of the previous two steps.   More experimentation can be and will be explored in the next experiments in the following fronts:   * Class Imbalance dealing: Dealing with class imbalance using various oversampling or under sampling techniques could help bridge the gap between the f1 scores of the majority and minority classes. * Deeper Machine learning models: Deeper and more comprehensive machine learning models can be explored into. This could help uncover and extract deeper underlying patterns in the dataset. * Feature Engineering: More feature engineering can be performed to extract informative features from the dataset. |
| **4.b. Suggestions / Recommendations** | Given the results achieved and the overall objective of the project, list the potential next steps and experiments. For each of them assess the expected uplift or gains and rank them accordingly. If the experiment achieved the required outcome for the business, recommend the steps to deploy this solution into production.  The potential next steps and experiments and go ahead with the project are as follows:   * Class Imbalance handling: Dealing with the existing class imbalance would be step 1 in the coming experiments. We see a steep rise in achieved scores just by implementing a deeper model. Bridging the gap between the majority and minority class would help the model learn and apply relevant information on both target variables. * Deeper models: Exploring into deeper and more comprehensive models like random forest, Gradient Booster etc. could help in uncovering further patterns in the dataset. * Feature Engineering: Investing time to explore into features in the existing dataset and extra features from external sources could help add perspective to the model.   Future steps to be taken with the project could be as follows:   * Privacy and Ethics: Ensure strict guidelines are established to ensure privacy and ethical concerns. The project relates to customer information. This should be done in a consensual and secure manner with proper established guidelines. * Inclusion: Ensure all groups of individuals are accounted for and represented. This includes people of aboriginal backgrounds, international immigrants, etc. * Aboriginal culture: As a company based in Australia, ensure Aboriginal culture, traditions, and customs are accounted for before deployment. This needs to be a separate sect of the company solely focused on ensuring all different nations and clans are accounted for in making decisions. This could include steps like language inclusion, cultural considerations etc.   The experiment requires more work before deployment can be thought of. The model still misses a high number of positive classifications which could indicate missing potential customers. |