# **EXPERIMENT REPORT**

| **Student Name** | Tarun Krishnan |
| --- | --- |
| **Project Name** | Assignment 2, Classification |
| **Date** |  |
| **Deliverables** | Part\_B.ipynb  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** | The business in question is a car dealership that wishes to analyse the dataset and produce information that will help sell cars.  The results of the model trained will be used to suggest which current customers and car owners are more or less likely to buy a new car.  Inaccurate results will result in customers that would actually buy cars not being approached and thereby losing out on sales, while accurate results will have the opposite effect. | |
| **1.b. Hypothesis** | I believe that there is a correlation between the various fields and the likelihood of buying a new or second car, and to that extent will train a SVM classification model that will help the business make better sales and further progress. | |
| **1.c. Experiment Objective** | The expected outcome here is to successfully identify the fields and characteristics that correlate to the likelihood of buying a new car and thereafter train a SVM classification model to help classify the category that people are to fall into to recommend the business to pursue them as favourable clients or not. | |

| 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** | The dataset was large, with over 100,000 entries. While most of the columns contained usable values, the age band and gender fields had a large number of NaN fields that needed to be cleaned.  Additionally, the categorical fields for car model and segment too were converted to numerical values to prevent overfitting.  Since the data exploration stage yielded the fields of use, the cleaning of all values may not have been required, however, for the sake of future use the cleaned code was used regardless. | |
| **2.b. Feature Engineering** | While no features were generated, when it came to train the model a bunch of features were dropped due to little/no correlation.  Features such as ID, age\_band and gender might have shown correlation, however, inspecting the dataset shows a large factor to this is bias introduced by sample size, which would not be truly or accurately indicative of the real world or of minorities. | |
| **2.c. Modelling** | The model used here is a Support Vector Machine Classifier. The strength of this model lies in the fact that it is fairly performant at segregating the data graphically.  The hyperparameters trained were:  kernel : Changes the transformation of the data to better separate the dataset  C : Changes the linearity of the decision boundaries  gamma : Changes the impact of the support vectors  class\_weight : Remove imbalance/bias due to overwhelming sample weight | |

| 1. **EXPERIMENT RESULTS** | | |
| --- | --- | --- |
| 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** | Training Data :  Accuracy Score : 0.9623104816834651  F1 Score : 0.9695757161471467  Testing Data :  Accuracy Score : 0.9568296025582458  F1 Score : 0.9653626553065908  Confusion Matrix for Training :  [[98281 3927]  [ 33 2828]]  Confusion Matrix for Testing :  [[24575 1033]  [ 101 559]] | |
| **3.b. Business Impact** | The confusion matrices are of the following format [[0, 1],[1, 0]]. While the accuracy and f1 scores seem to be high, the confusion matrix actually allows us to understand why.  The model is fairly accurate predicting true negatives and true positives. The issue lies with the false positives and more importantly false negatives.  As the business model strongly relies on making sales on cars rather than saving money by not trying to sell a car, it is better in this business to predict fewer false negatives and make more predictions that are false positives.  After all, it is better to lose money by being unable to sell a car to an unwilling customer than to lose money by not wanting to sell a car to a willing customer. | |
| **3.c. Encountered Issues** | The only visible issue with this dataset and training the model stems from the overwhelming class disproportion and the size of the dataset that heavily reduces the ability to effectively train models due to bias and time complexity required to actually tune the hyperparameters. | |

| 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** | The outcome of this experiment seems rather satisfactory. Given the dataset and the tuning performed, it has performed up to expectation. Different models could give different results, so while this particular experiment has concluded, further experiments can improve upon this. | |
| **4.b. Suggestions / Recommendations** | While this experiment may be over, we could try to train other models and make better predictions based off of those if we can indeed improve our accuracy and precision.  That being said, this model is slightly overfit to the training data and shows poorer performance compared to logistic regression, if only marginally so. | |