# **EXPERIMENT REPORT**

| **Student Name** | Tarun Krishnan |
| --- | --- |
| **Project Name** | Assignment 2, Classification |
| **Date** |  |
| **Deliverables** | Part\_F.ipynb  K Nearest Neighbour Classifier |

| 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 K Nearest Neighbour Classifier 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 K Nearest Neighbour Classifier 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 trained for this experiment was a K Nearest Neighbour Classifier. Inherently, it tries to find the class a datapoint belongs to by finding its closest neighbours.  While there are hyperparameters, tuning them yielded no actual improvement or change, and so they were discarded.  Instead, after trial and error, the number of neighbours for an optimal result was found to be 3. | |

| 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 score : 0.9905395502003446  Testing score : 0.9862951119232526  Confusion Matrix for Training :  [[102081 127]  [ 867 1994]]  Confusion Matrix for Testing :  [[25544 64]  [ 296 364]] | |
| **3.b. Business Impact** | While the accuracy scores seem to be high, the number of false negatives is a strong deterrent from using this in an actual deployment.  The models trained up till now show better results and are all viable compared to the negative impact that this model’s performance would have due to the presence of false negatives. | |
| **3.c. Encountered Issues** | Other than the overwhelming class disproportion and the size of the dataset that heavily reduces the ability to effectively train models, we can not make any improvements through the hyperparameters, and so need to re-evaluate the use of this model going forward. | |

| 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 unsatisfactory. Given the use ase of the business, it would be wise to pursue other models. While a more consolidated dataset could provide better insight, the primary take away from this is that the points are disorganised and unarranged to use with this algorithm. | |
| **4.b. Suggestions / Recommendations** | While this experiment may be over, we could further investigate and try to manually tune the model to fix issues stemming from underfitting and randomisation. | |