# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Tahmidul Islam |
| **Project Name** | Logistic Regression |
| **Date** | 01-05-2023 |
| **Deliverables** | Attached File: Experiment 1 – Logistic Regression.ipynb  Model: Logistic Regression with L1 and L2 Regression and KNN |

|  |  |
| --- | --- |
| 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 goal is to identify the best classification model that can predict the likelihood of an individual buying another car.  The accuracy scores will be evaluated, and depending on how the model performs on the training, validation and testing set, the effectiveness of the model will be judged.  The accuracy or correctness of results will influence marketing strategies and business decisions that want to determine if a person having a certain model of car driving a certain mile of distance, etc. will buy another car or not. |
| **1.b. Hypothesis** | The question we want to answer is if there is a relationship between a customer’s current car model, vehicle type, payments in services, etc. with the likelihood of him buying a second car.  It is worthwhile considering it because the current situation of a car can often determine a person’s decision to buy the next. For example if a person has an old, worn out model of a car that is costing him a lot of money on maintenance, it is highly likely that he would purchase a better car. In addition, there may be car fanatics out there who just likes to experiment with different cars, and there are often patterns to it. Thus our project focuses on identifying the type of person the customer is based on his car conditions and make a prediction accordingly. |
| **1.c. Experiment Objective** | The expected outcome is that the model we are building will have an accuracy score that will help us to identify the strength of the relationship (as stated on our hypothesis).  The score could be very high, indicating that the model is very successful in its predictions. However, if it is very high it means the model is overfitting and thus needs to be adjusted.  The score could be lower indicating the opposite. |

|  |  |
| --- | --- |
| 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** | Before data preparation, a copy of the dataset was made and all the modifications were made on them so that the actual copy is preserved. Some columns were dropped because of having too many null values. Then we encoded the columns that were categorical. Once done, we found out that half the features of the dataset had 0 and 1 values while the other half had values that ranged upto 10. Thus we normalized the data to ground all the values.  Once all was done, we split the data into train, validation and test variables.  For future experiments, data scaling may be potentially important to ensure that values of only a handful of features get to express their significance in the model. |
| **2.b. Feature Engineering** | In the dataset, we have the ‘car\_model’ as a categorical value. Now the issue is that we can either consider it as an ordinal by assuming the model names refer to some form of chronology, and higher models mean newer brands. Or we can consider it as a nominal which means that the model names cannot be placed in any sequence. I have assumed it as nominal because it could be possible that model\_15 and model\_16 for example, have been introduced in the same year for that car brand. Thus in this case, we cannot put model\_15 and model\_16 in any sequential form.  Consequently, One-hot encoding was performed on the column which produced 19 more columns (or features) of boolean values.  We also had to remove some columns. We found out that there were a high number of missing values (85%) of the age\_band column and the gender (52%) column.  An initial thought was to apply mode imputation to the age\_band column, but since it would cover up 85% of the values, making such an assumption would be risky. Thus I decided to drop the column. Consequently, this will act as one of the limitations of the project that age being such a seemingly important factor was not considered, but this is for the sake of training our model correctly.  The same decision was made for the gender column since half the data was missing.  The features car\_segment (size) and ag\_of\_vehicle\_years might be important for future experiments as it has the highest chances of determining the decision of vehicle purchase. |
| **2.c. Modelling** | The model trained was Logistic Regression with L1 and L2 Regression and KNN and it was chosen for its simplicity and ease of implementation. In addition, it works well with large datasets and can be regularized to prevent over-fitting and improve performance of the algorithm.  The models that were not used were SVMs, Decision Trees and Random Forests because they are too complex for now. However, they maybe implemented in future experiments.  An important hyperparameter for the logistic regression model is the penalty parameter. |

|  |  |
| --- | --- |
| 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** | Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes.  Accuracy of:  baseline model = .97349  training set = 0.97782  validation set = 0.97891  applying regularisation on train = 0.97783  applying regularisation on validation = 0.97892  knn 15 neighbors on train = 0.983534  knn 15 neighbors on validation = 0.982393  knn 55 neighbors on train = 0.97788  knn 55 neighbors on test = 0.97704  With Logistic Regression, for some reason we see a higher accuracy on the validation set than on the training set. This is a rare situation and happens plausibly because of overfitting. This problem persists even after using L1 and L2 Regularization, suggesting that some other hyperparameters need to be changed.   However with KNN 15, the problem is not present. We have an increased accuracy with this model. This could mean that the relationship of the target variable with the features is slightly less linear and more complex than we thought at the beginning, and KNN could capture it. |
| **3.b. Business Impact** | The results of the experiments tell us that the decision to buy a second car is indeed related to the parameters like age of car, model of car and expenses behind maintaining the car. Incorrect results may have serious level impacts on business as it will suggest the wrong business strategy to be implemented. |
| **3.c. Encountered Issues** | One of the major issues during the experiment is the age\_band and gender column as mentioned before. The workaround was to drop these columns entirely, but this may not be a good solution as these two were very important parameters. I wanted to mode imputations to the rows, but then I would run the risk of making assumptions for 85% of the data. This issue may persist in future experiments. |

|  |  |
| --- | --- |
| 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 confirms our hypothesis pretty well since we have obtained a very high accuracy score on the test dataset. However, there were cases of overfitting found and regularization failed to mitigate it. We thus may need to try out a few more classification algorithms and find out if they can help. |
| **4.b. Suggestions / Recommendations** | The next steps of the project may be to re-define the dataset by obtaining values for the age\_band and gender parameters. Model-wise, the experiment is doing well but we just need to include the mentioned features to be sure that we have a strong model. This model can be deployed into production. This can be done by Containerizing the application like in Docker and choosing an appropriate infrastructure for hosting the application like AWS or GCP. |