# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Tahmidul Islam |
| **Project Name** | Classification with Random Forest |
| **Date** | 01-05-2023 |
| **Deliverables** | Attached File - Experiment 4: Binary Classification with Random Forest.ipynb  Model: Random Forest |

|  |  |
| --- | --- |
| 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 if Random Forest classification model can successfully 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 them 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** | For 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 Random Forest whose advantages over other classification algorithms like RANDOM FOREST is speed. It also provides information about the importance of input features.  The models that are not used here are Linear Regression, SVM and Decision Trees. However, they are implemented in separate experiments.  An important hyperparameter for the Random Forest model is n\_estimators which denotes the number of decision trees in the forest. Increasing trees usually improves accuracy, but comes with additional time and space complexity. |

|  |  |
| --- | --- |
| 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.  Default RANDOM FOREST on validation set:  F1 Score on Validation: 0.99248  RANDOM FOREST with n\_estimators = 50 on validation set: F1 Score = 0.99243  RANDOM FOREST with n\_estimators = 200 on validation set:  F1 Score = 0.99238  RANDOM FOREST with n\_estimators=50, max\_depth=15, min\_samples\_leaf=2, max\_features = 5 on validation set:  F1 Score = 0.99129  RANDOM FOREST model with n\_estimators=50, max\_depth=15, min\_samples\_leaf=10 on validation set   |  |  |  | | --- | --- | --- | |  | Predicted Label [0,1] | | | True Label [0, 1] | 1 | 0.00078 | | 0.41 | 0.59 |   RANDOM FOREST model with n\_estimators=50, max\_depth=15, min\_samples\_leaf=10 on test set   |  |  |  | | --- | --- | --- | |  | Predicted Label [0,1] | | | True Label [0, 1] | 1 | 0.00071 | | 0.41 | 0.59 |   We first run a default model of Random Forest after which we tune the parameter by increasing n\_estimators to 50. But doing so doesn’t give us much change on accuracy scores. (Accuracy score is already very high though). However we increase the parameter and then add some more parameters like min\_samples leaf and max\_features but still see no further change.  Looking at the confusion matrix we see that the False Negative is high. This means that the model performs less well in finding out negative results. |
| **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 matrix on the test dataset. However, our False Negatives are not very low, meaning Recall is low.  Doing this project gives us a good insight into feature importance. From the report, we can see that car\_segment feature is the most important one in influencing the decision to buy another car. The next 2 in line are age\_of\_vehicle\_years and sched\_serv\_warr. Thus for the business, it may be helpful to mostly focus on the type of vehicle that a customer has, the age and the warranty of that vehicle so that they may make easy, rudimentary estimates with their predictions.  At this point, I think sufficient insights could be obtained from the experiment and previous experiments. Just to venture out some more, we can use this model, but getting the optimum hyperparameter first, using some algorithms. This is something that can be done in a further experiment. |
| **4.b. Suggestions / Recommendations** | The project’s next steps may be to re-define the dataset by obtaining values for the age\_band and gender parameters. Model-wise, the experiment is doing very good. 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. |