J]UJWNR JSY%WJUTWY

| | |
|---|---|
| **Student Name** | Shalimar Chalhoub |
| **Project Name** | Classification Models |
| **Date** | 25/3/2023 |
| **Deliverables** | <MLAA - Assignment 2 Method 5 - Random Forest with Automated Hyperparameter tuning><br><Radom Forest with Automated Hyperparameter Tuning> |

---

| **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 of this project to the business is to accurately identify which customers are going to be repurchasing a second car from the company depending on several factors such as their gender, age and the model and segment of their car as well as some other details about the car.<br>These results can be used to target those types of customers and spend more resources on getting them to buy a second car.<br>Inaccurate results may cause the company to lose money by targeting the wrong customers as well as lose sales. |
| **1.b. Hypothesis** | I want to test whether Random forest with automated hyperparameter tuning can perform better than when I manually chose the parameters<br><br>The reason behind me choosing random forest is because it usually has a high accuracy because the decision is being taken by many trees instead of one and it works well with noise and outliers and can handle big non-linear datasets such as the one I have. It also has feature importance, so it measures the importance for every feature of the dataset which helps the interpretability of the model<br>The reason why I decided to do automated hyperparameters is because I want to check multiple combinations and find the one that suites my model best without having to keep testing |
| **1.c. Experiment Objective** | I think the outcome of the random forest with automated tuning should be better than when I chose my own parameters in the previous experiment so a recall above 73%<br><br>Possible outcomes:<br>1. The model can either have a large recall and thus be efficient to be deployed by the company<br>2. Or, the model with produce a low recall score and would be dismissed |

| 2. 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 this model, I have used datasets that have already been preprocessed and split during the first experiment and thus no data preparation took place.<br>The rationale has been explained in the first experiment report<br>The steps I decided not to use is data balancing, as I have tried it and it made my model's predictions uneven and skewed. |
| **2.b. Feature Engineering** | Feature engineering was performed before splitting the datasets in experiment 1.<br>It included imputation of gender and age_band variables as well as checking whether a model had 2 different car_segments and fixing them<br>It was found that model_17 had the segments small/medium as well as others and thus others was transformed to small/medium for that model |
| **2.c. Modelling** | The model used for this experiment is RandomForestClassifier which is an ensemble method that combines multiple decision trees which will be trained on a subset of the data and features and the best prediction will be chosen by majority voting. I also combined it with GridSearchCV to find the best parameters<br>Best hyperparams:<br>1. n_estimators =30<br>2. Max_depth=25<br>3. min_samples_leaf=10<br>I also tested RandomizedSearchCV with Kfold however, I decided not to use it as GridsearchCV produced better results<br><br>I decided not to sample for this experiment as I found it gave me biased results, however, for the future we can try other sampling techniques |

| 3. 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** | The recall scores are listed below:<br>Train set:0.64<br>Dev set:0.58<br>Test set: 0.59<br><br>The results are low compared to the other models that we have used and that is because the dataset is very skewed and imbalanced which then makes the parameters choosing biased as well |
| **3.b. Business Impact** | I think the results are low and do not suite the business needs and we can definitely go with one of the previous models over this one as deploying it might cause the business to lose money by not marketing the correct people. |
| **3.c. Encountered Issues** | These are the issues encountered with this experiment:<br>1. Missing data: this was fixed at the beginning of experiment 1 using imputation<br>2. Imbalanced data: this has not been fixed as the model performs well without it<br>3. low results: this I wasn't able to fix because it was automated hyperparameters |

| 4. 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** | I think overall, this experiment did not do as well as I hoped it would because the parameters chosen were not the best. I did learn that imbalanced data need a lot of work and fixing in order to predict good results<br><br>I think this project is not a dead end however, there might be a need to fix the issue of the imbalanced dataset using a method that will not affect the biasing upon prediction |
| **4.b. Suggestions / Recommendations** | I think even though this model did not reach the desired above 75% outcome, it can still be usable if it is the only model available because even though the recall data is low, the precision score is high, meaning that by marketing everyone on the list, even if we are not reaching all the desired clients, at least we are not reaching a lot of wrong ones and thus would not be a total loss |