# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | **Savinay Singh** |
| **Project Name** | A comparative analysis of different classifiers for predicting the likelihood of an existing customer buying a new car |
| **Date** | 27 April 2023 |
| **Deliverables** | Experiment3.ipynb  SMOTE  Cramer’s V  Feature Scaling  Logistic Regression  XGBoost  Naïve Bayes  K Nearest Neighbours  Support Vector Machine  Random Forest  Seaborn, Matplotlib  Exploratory Data Analysis |

|  |  |
| --- | --- |
| 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** | Explain clearly what is the goal of this project for the business. How will the results be used? What will be the impact of accurate or incorrect results?  The main objective of this project is to compare the performance of different classifiers to predict the likelihood of an existing customer buying a new car.  The outcome can be useful for the car companies to:   * + - 1. Develop and execute retention tactics that specifically target customers who are likely to repurchase the car.       2. Better inventory management as the companies would know what models have a higher chance of being resold.       3. Enhancing customer satisfaction by improving certain aspects of car service.   The accurate prediction will give car companies crucial knowledge to focus on the areas that need improvement. They can stay ahead of the competition and improve their overall business performance. |
| **1.b. Hypothesis** | Present the hypothesis you want to test, the question you want to answer or the insight you are seeking. Explain the reasons why you think it is worthwhile considering it,  The hypothesis proposes that by taking testing and training accuracy as a metric, a modelling technique can be finalized for predicting the probability of an existing customer purchasing a new car in an efficient manner. This model can help car companies to execute retention tactics that specifically target customers who are likely to repurchase the car. The companies can focus on the features that can help them increase their customer retention rate. |
| **1.c. Experiment Objective** | Detail what will be the expected outcome of the experiment. If possible, estimate the goal you are expecting. List the possible scenarios resulting from this experiment.  The expected outcome of this experiment is to get a most accurate model by performing a comparative analysis of different classifiers for predicting the likelihood of an existing customer buying a new car. This can help dealerships and manufacturers better understand the factors that influence customer behavior and decision-making, thereby allowing them to improve their services and offerings accordingly. The objective is to attain an accuracy rate higher than 90%, but since there is an issue with uneven class distribution, the model will be considered reliable if it generates results with satisfactory precision and recall values for both positive and negative categories. |

|  |  |
| --- | --- |
| 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** | Describe the steps taken for preparing the data (if any). Explain the rationale why you had to perform these steps. List also the steps you decided to not execute and the reasoning behind it. Highlight any step that may potentially be important for future experiments  Data preparation is done by the following steps:   * + - 1. *Handling the missing data*: It is important in predicting the likelihood of an existing customer buying a new car because it can bias the results of the analysis and reduce the accuracy of the model.       2. *Label encoding* of the gender, car model and car segment to convert textual labels into numerical categories.       3. *Handling the class imbalance*: SMOTE is a technique used here to generate synthetic samples for the minority class by interpolating between neighboring examples. This helps to balance the distribution of classes and improve the performance of the model.   Steps decided not to be executed:   * *One-hot encoding all decile categories, regardless of their cardinality*: This step could potentially lead to the curse of dimensionality because the number of features increases drastically, and the model becomes harder to train and interpret.   The steps that can be taken during data preparation and are expected to improve the outcome in the future are:  *Synthetic data generation*: Creating new data points that are statistically similar to the existing dataset in order to expand the horizon of the project.  *Check for multicollinearity*: The model may experience issues due to multicollinearity, including inaccurate results from hypothesis testing. Thus, it is crucial to examine a regression model's multicollinearity. |
| **2.b. Feature Engineering** | Describe the steps taken for generating features (if any). Explain the rationale why you had to perform these steps. List also the feature you decided to remove and the reasoning behind it. Highlight any feature that may potentially be important for future experiments   * + - 1. Removal of a feature “age\_band” which had more than 85% of missing values.       2. Removal of ID because it is a unique identifier for each customer and is not a relevant feature       3. Filling the null values of Gender column by “third gender/Not stated”.       4. Label encoding of the gender, car model and car segment to convert textual labels into numerical categories.       5. Feature scaling to bring all the features of the dataset to a similar scale or range of values.       6. The number of months since the last service is the most important feature for prediction for random forest classifier in this case because of its high feature importance score.     No new feature was generated out of the existing features in this experiment. |
| **2.c. Modelling** | Describe the model(s) trained for this experiment and why you choose them. List the hyperparameter tuned and the values tested and also the rationale why you choose them. List also the models you decided to not train and the reasoning behind it. Highlight any model or hyperparameter that may potentially be important for future experiments  This study employs use of multiple classical machine learning models such as Logistic Regression, XGBoost, Naïve Bayes, K Nearest Neighbours, Support Vector Machine, Random Forest. The accuracy of the training and testing data is then compared to determine the model with the highest performance. Also, the Cramer’s V value for each feature on the basis of target variable is less than 0.5.    Random forest classifier gave the highest accuracy and has the following hyper parameters:  {'bootstrap': True,  'ccp\_alpha': 0.0,  'class\_weight': None,  'criterion': 'gini',  'max\_depth': None,  'max\_features': 'sqrt',  'max\_leaf\_nodes': None,  'max\_samples': None,  'min\_impurity\_decrease': 0.0,  'min\_samples\_leaf': 1,  'min\_samples\_split': 2,  'min\_weight\_fraction\_leaf': 0.0,  'n\_estimators': 100,  'n\_jobs': None,  'oob\_score': False,  'random\_state': None,  'verbose': 0,  'warm\_start': False} |

|  |  |
| --- | --- |
| 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.  The retention rate was predicted with 99% accuracy using the Random Forest classifier. The experiment also demonstrates that paying a lower amount for scheduled service increases the probability of a person buying the car again and is proved to be most significant factor.    The precision, recall and f1 score are significantly high for both the positive and negative classes. |
| **3.b. Business Impact** | Interpret the results of the experiments related to the business objective set earlier. Estimate the impacts of the incorrect results for the business (some results may have more impact compared to others)  The estimated outcome of this experiment is to build a model that can help dealerships and manufacturers better understand how the different features of the automobile can impact the likelihood of an existing customer buying a new car. The experiment also demonstrates that paying a lower amount for scheduled service increases the probability of a person buying the car again and is proved to be most significant factor. The precision, recall and f1 score are significantly high for both the positive and negative classes. |
| **3.c. Encountered Issues** | List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them. Highlight also the issues that may have to be dealt with in future experiments.  Although there is no mention of any sensitive information such as vehicle identification number (VIN) or Personal Identifiable Information (PII) in the dataset, there could still be some data privacy threats such as:   * + - 1. Repurchase data can be utilised for targeted marketing initiatives, which some customers may find intrusive or unwanted. Customers' privacy may have been violated if their information was utilised for marketing without their explicit permission.   **Data Quality Issues**   * + - 1. Small Dataset: The dataset might not be representative of the entire population, so there is a higher chance of bias in the model.       2. Imbalanced dataset: Imbalanced dataset may be biased towards the majority class (i.e. Customer not buying the car again), resulting in poor performance on the minority class (Customer buying the car again).       3. High missing values: Some features such as age\_band had really high volume of missing values and had to be removed as they can potentially lead to bias while training the model.     **Modelling Issues**   1. Risk of overfitting: Due to the small size of the training data there was a huge risk of model overfitting. 2. Features not showing a high correlation: The features selected in the study were not showing a high correlation with the target outcome. |

|  |  |
| --- | --- |
| 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** | Reflect on the outcome of the experiment and list the new insights you gained from it. Provide rationale for pursuing more experimentation with the current approach or call out if you think it is a dead end.  The experiment was successful with high f1 score for both positive and negative classes, the model is able to accurately predict the likelihood of an existing customer buying a new car based on their highly correlated features. This could be useful for marketing campaigns, customer retention strategies, and inventory management in the automotive industry. It is possible that the high accuracy is a result of the imbalanced nature of the dataset and limited data available for the positive class. Therefore, including more features and data in the model can improve its reliability and mitigate the effects of class imbalance. |
| **4.b. Suggestions / Recommendations** | Given the results achieved and the overall objective of the project, list the potential next steps and experiments. For each of them assess the expected uplift or gains and rank them accordingly. If the experiment achieved the required outcome for the business, recommend the steps to deploy this solution into production.  The potential next step would be to fine-tune the hyper-parameters of the random forest model.  There are many ethical issues also involved while applying any machine learning model to a complex task such as:   * + - * Bias in the training set       * Transparency and Explainability of the model       * Privacy and Security concerns       * Accountability and responsibility   So, the potential next steps will also involve taking these into consideration.  Deploying the final model and making it available for use in production can be done using the following potential steps:  Exporting the model in a pickle file.  Preparation of the data pipeline. CI/CD (Continuous Integration and Continuous Deployment) pipeline is better suited as it focuses on building, testing, and deploying software quickly and reliably. The pipeline also reduces the possibility of bugs and errors entering the production environment.  Deploying the model on a cloud-based platform or a web server. Then, planning the architecture of the containerization and orchestration of the application.  Testing if the model is performing correctly in the production environment.  Monitoring the results of the model over time. This also involves creating a dashboard to keep track of the model's performance and detect any issues.  Improving the model by re-training it over a period of time on the new and updated data. |