# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Ronik Jayakumar |
| **Project Name** | MLAA Assessment 2 Experiment 3 |
| **Date** | 25/04/2024 |
| **Deliverables** | Experiment 3  Random Forest 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** | 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?  We are working for a car reseller company who is planning their next marketing campaign. The goal of the described project is to build a model that can accurately predict what segment of customers are most likely to buy another car so that marketing efforts and resources can be efficiently used on the right customer category.  The results of the project will be used to identify customers who fall under a defined category based on certain business questions asked and existing data on customers. The model will be run on this data to identify customers of a certain category who are most likely to make another car purchase so that marketing efforts can be used on them.  The impact of accurate results would be correct customer segmentation and target group identification which would result in high sales turnover.  On contradiction, the impact of inaccurate results would result in wrongful customer identification which would allow marketing resources to be expelled on customers not likely to make purchase, thereby reducing profitability. |
| **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 that is being tested is to check if past trends in car maintenance can predict the correlation with a customer’s inclination to buy a second car.  The insights required relate to what are the most useful questions to put forth to customers that would help us accurately find the customer segment most inclined towards investment into a new car.  It is worthwhile considering this hypothesis and investing into this model as it would help reseller company make informed data driven decisions. A company could have hundreds of thousands of potential customers on their database. Identifying the right customer segment so that marketing efforts are rightly targeted is key.  Having an accurately working model can ensure right segments are being targeted which in turn would increase sales turnover and overall revenue. |
| **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 a model that can detect, classify, and predict a certain segment of customers, based on past car related activities, on if they are likely to invest into a second car.  An estimation of the expected goal would be a model that achieves at least an 80% overall accuracy in ensuring the right customer group is identified.  The possible scenarios resulting from this experiment are as follows:   * Successful Prediction – The model identifies the right customer, and the customer makes a purchase. * Missed Opportunity – The model fails to identify a potential customer, and the customer ends up purchasing a car. * Successful Rejection – The model correctly identifies an unlikely to purchase customer. * False rejection – Model identifies a wrong customer as someone who is likely to purchase a car which results in wasted marketing efforts. |

|  |  |
| --- | --- |
| 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 steps have been imported from experiment 1. The cleaned and prepared dataset was downloaded at the end of experiment 1 and this dataset will be used for the remaining experiments.  The steps chosen to not be executed is in changing the features chosen to go ahead with Machine Learning, one of the things kept in mind during the experimentation phase is to ensure all features remain constant throughout so an accurate reading on the differences between various models and data handling techniques affect the final output.  Steps that could be potentially important for future experiments:   * Collect and deploy more data on customers who made a purchase of a second car. The existing data has a high class imbalance between customers who did not make a purchase and the customers who did, corresponding to the 0 and 1 values in the target column. * More inclusion: Ensure more inclusion in data collection techniques to ensure representation of people from all aboriginal cultural fronts. |
| **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  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.  Feature Engineering has been performed in the dataset.   * **Gender Column**: The column contained approximately 50% of all the values as null with the other 50% corresponding to male or female. Since a high number of rows were null, dropping them would mean getting rid of a large amount of useful data. This has been tackled using Multivariate Imputation by Chained Equations or MICE for short. The following steps were taken.   + The entire column was first encoded, followed by which the Nan class was reverted to its original form to make sure null values get predicted.   + The imputing was done into a new column gender column where all the null values were predicted.   + The values were mapped back to their base form, that is as Male and Female. * **Feature Selection**: Feature selection techniques have been performed to decide what features have the highest relation to the target column. The car\_segment and the gender\_no columns are chosen to ensure representation of people and the type of car. Along with this the following was performed:   + **Correlation Analysis:** A correlation matrix was calculated between all the columns and the target column. The correlation of each column with the target column was visualized using a heatmap.   + **ANOVA-F test**: The ANOVA test was carried out as well as a second check on the correlation analysis. Once the features with the highest influence were identified, they were isolated and a new DataFrame was created with these features along with the 2 chosen categorical features. * **Class Imbalance:** The high class imbalance that exists in the target column has been dealt with in experiment 2 using ADASYN oversampling technique. It focuses on the minority instances that are difficult to classify correctly, rather than oversampling all minority instances uniformly. It assigns a different weight to each minority instance based on its level of difficulty in classification. [1] |
| **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.  Baseline + Main model: Random Forest Classifier:  The random forest classifier has been used as both the baseline and the main model in the experiment to compute the difference between the model being deployed on the original dataset available and the model being deployed on the dataset after the class imbalance has been dealt with.  The random forest classifier was chosen to experiment with due to the following reasons:   * **Diversity:**Not all attributes/variables/features are considered while making an individual tree; each tree is different. * **Immune to the curse of dimensionality:** Since each tree does not consider all the features, the feature space is reduced. * **Parallelization:**Each tree is created independently out of different data and attributes. This means we can fully use the CPU to build random forests. * **Train-Test split:**In a random forest, we don’t have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree. * **Stability:**Stability arises because the result is based on majority voting/ averaging.   Hyperparameters tuned:   * n\_estimators: The total number of trees in the forest. Increasing this number may enhance model performance by lowering variance, but it raises computing costs. * max\_depth: The greatest depth of any tree in the forest. Controlling tree depth helps to avoid overfitting. A deeper tree can catch more complicated patterns, but it is susceptible to overfitting. * min\_samples\_split: The smallest number of samples needed to split an internal node. Increasing this amount can help prevent overfitting by requiring more samples per split. * min\_samples\_leaf: The minimal number of samples needed to reach a leaf node. Increasing this amount can help prevent overfitting by requiring a minimum number of samples in each leaf. * Max\_features: The number of features to examine when determining the optimum split. 'auto' utilises all features, whereas 'sqrt' takes the square root of the entire number of features. Limiting the number of features can decrease overfitting and increase computing efficiency. * Bootstrap: Whether to bootstrap samples while creating trees. Bootstrapping adds unpredictability and variation to each tree, which can help model performance.   For future experiments, one hyperparameter that may be particularly important to explore further into is max\_depth. The max\_depth parameter controls the maximum depth of each tree in the forest. It plays a crucial role in controlling the balance between model complexity and overfitting. |

|  |  |
| --- | --- |
| 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 performance metrics for the Random Forest classifier are as follows:    As we can see, the model performs extremely well with the training, validation, and test data for both the majority and minority features.  The main underperforming features were as follows:   * Minority class prediction before class imbalance: Minority class predictions worked extremely well with the Random Forest classifier only after class imbalance was handled. * Model Optimization: More model optimization is required to ensure little to none of the class 1 target variables are missed. As we can see with this business structure, the turnover rate is low. Hence missing potential customers could prove a costly mistake to the company. |
| **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)   * Accuracy: The model's high overall accuracy (97% on both validation and testing sets) indicates that it is capable of properly predicting both classes (0 and 1) most of the time. This shows that the model can give useful information on automobile repurchase behavior. * Precision and Recall: While class 1 (minority class) has reasonably good precision and recall (96% and 98%, respectively), there is still potential for development, especially in precision. Misclassifications of class 1 incidents may result in missed chances to target potential automobile repurchasers, thus costing sales or income. * F1-Score: The F1-score, which is the harmonic mean of accuracy and recall, offers a fair assessment of the model's performance. With an F1-score of 0.97 for class 1 on both validation and testing sets, the model performs well in detecting cases of automobile repurchase. * Cost Benefit Analysis: Performing a cost-benefit analysis can assist assess the possible impact of applying the categorization model in real-world circumstances. * Decision Support: The categorization model may be an effective decision-making tool for company stakeholders such as marketing teams and sales departments. The model, which successfully predicts auto repurchase behavior, can assist in prioritizing resources and customize marketing efforts to target consumers most likely to repurchase.   The impact of incorrect results could indicate loss in customers due to missed marketing opportunities thereby indicating loss in revenue. |
| **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.  Problems faced during experimentation are as follows:   * Minority class sampling: The class imbalance was one of the issues faced in the experimentation phase. There exists a massive gap between the 0 and 1 target variables. * Limited features: The dataset consists of a limited number of features. This could be a positive in getting more customers to respond to surveys, but the overall number of features that the model gets to learn stays highly limited.   Future problems:   * Future problems could arise in the following ways:   + Privacy issues: Data collection on car maintenance should comply with privacy laws and regulations. Privacy, data storage, and data management could a problem. * Underrepresentation of Aboriginal culture: The aboriginal culture is a wide array of multiple nations and clans with different cultural beliefs and aspects. Ensuring proper representation and consideration into the collection, processing, and storage of data needs to align with the custodians of the land. Ensuring proper research and groundwork is done before going ahead with the project would be avid. |

|  |  |
| --- | --- |
| 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.   * High Accuracy and balanced metrics: The Random Forest model performed well in terms of accuracy and balance, with precision, recall, and F1-score all either 0.96 or above for both classes (0 and 1) on both the validation and testing sets. This suggests that the model is effective in discriminating between automobile repurchasers and non-repurchasers. * Potential or further optimization: While the random forest model worked admirably overall, there may be opportunity for more optimization. Experimentation with hyperparameter tweaking, feature engineering, or ensemble approaches may improve the model's performance even more and reveal new insights regarding automobile repurchase behavior. * Cost Effectiveness: Decision trees are computationally efficient and have shorter training and prediction durations than more complicated models such as neural networks or gradient boosting machines. This makes them an economical solution for modelling jobs, especially in circumstances where interpretability is critical.   Potential Dead Ends in the Experiment could include the fact that while decision trees have many advantages, their effectiveness may be limited, especially when dealing with complicated links and interactions in data. If additional decision tree testing does not dramatically enhance model performance or provide new insights, it may be worthwhile to examine different modelling techniques or rethink the issue formulation. |
| **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.  Potential Next Steps with the experiments are as follows:   * Different models: Exploring into deeper and more comprehensive models like Gradient Booster, Neural Networks etc. could help in uncovering further patterns in the dataset. * Feature Engineering: Investing time to explore into features in the existing dataset and extra features from external sources could help add perspective to the model. * Other Sampling techniques: We have used the ADASYN links oversampling technique in this experiment. Experimenting into other sampling techniques could provide insights into the overall model performance.   Future steps to be taken with the project could be as follows:   * Privacy and Ethics: Ensure strict guidelines are established to ensure privacy and ethical concerns. The project relates to customer information. This should be done in a consensual and secure manner with proper established guidelines. * Inclusion: Ensure all groups of individuals are accounted for and represented. This includes people of aboriginal backgrounds, international immigrants, etc. * Aboriginal culture: As a company based in Australia, ensure Aboriginal culture, traditions, and customs are accounted for before deployment. This needs to be a separate sect of the company solely focused on ensuring all different nations and clans are accounted for in making decisions. This could include steps like language inclusion, cultural considerations etc.   The model worked effectively and can be considered for deployment. The steps for deployment could be as follows:   * Integration with Production Environment: Integrate the model into the production environment by working with IT and DevOps teams to install it on existing systems and infrastructure. * Privacy and Ethical laws: Ensure privacy guidelines are established with ethical and cultural considerations especially towards underrepresented classes such as aboriginals. * Monitoring and Logging: Implement real-time monitoring and recording tools to keep track of model performance and spot abnormalities or deterioration quickly.   Testing and Validation: Thoroughly test and validate the deployed model to guarantee its dependability and resilience under a variety of scenarios and edge cases. |

References:

# [1] : Data Imbalance: How is ADASYN different from SMOTE? – Vijay M – Sep 26th 2023 - https://medium.com/@penpencil.blr/data-imbalance-how-is-adasyn-different-from-smote-f4eba54867ab