# **EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | **BALAKUMARAN SIVANESAN** |
| **Project Name** | Assignment 2: Classification models -Experiment 5 |
| **Date** | 28 April 2024 |
| **Deliverables** | <notebook nameLr.ipynb>  <model name: Lr>  <other> |

|  |  |
| --- | --- |
| 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 primary business objective of this predictive analytics project is to enhance the targeting efficiency of marketing campaigns by identifying customers who are likely to buy a new car. This focused strategy attempts to allocate resources more wisely toward high-probability leads in order to save marketing expenses and raise campaign conversion rates.  Furthermore, by enabling more customized interactions and offerings, the model's insights may greatly improve customer relationship management, which may increase customer happiness and loyalty. Precise forecasts are also helpful for inventory management since they help determine which automobile models are most likely to be in demand, which helps maintain optimal inventory levels and cut down on overstocking expenses. |
| **1.b. Hypothesis** | The hypothesis for this experiment is that certain customer behaviors and characteristics such as frequency of service engagement, preferred car features, and previous purchase history are strong predictors of whether a customer will repurchase a car.  Question: Can we predict a customer's decision to buy a new car based on their past interactions and preferences?  Reasons for considering the hypothesis:   * It's been shown that demographic information may forecast consumer behaviour in a variety of industries, including the car industry. * Any underlying patterns or correlations between gender/age and repurchase behaviour may be found by adding demographic data into the model.   - If adding demographic data improves the accuracy of the model, it could have a significant effect on customer retention and focused marketing strategies. |
| **1.c. Experiment Objective** | This study aims to investigate if car model, auto segment, and customer demographic data can improve the accuracy of predicting which customers are likely to repurchase a vehicle. The goal is to outperform the current baseline model's accuracy.  This experiment could lead to the following situations:  - If the new model does not outperform the baseline model, it might not be profitable using these additional characteristics to anticipate customer repurchase.  - If the new model only slightly outperforms the baseline model, then it might not be worth the additional efforts needed to gather and analyse the additional data.  - If the new model exceeds the baseline model, using customer demographics, vehicle models, and auto segment data could be a beneficial tool for projecting consumer repurchase and could potentially result in more sales and revenue for the company. |

|  |  |
| --- | --- |
| 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** | In the given code, the following steps were taken for preparing the data:  1. Drop irrelevant columns: The 'ID' and 'age band' columns were eliminated from the dataset as they were deemed redundant. These columns were removed since it was found they were not relevant to predicting the likelihood that a customer would make another purchase.  2. Encode category variables: The 'gender' column was encoded using Label Encoder, which converts categorical data into numeric values. This was necessary since machine learning algorithms typically work with numerical input, but categorical variables can only be included in the model by encoding them.  3. One-hot encoding of categorical variables: The 'car\_model' and 'car\_segment' columns were one-hot encoded using pd.get\_dummies. This was necessary since the variables were category-based and had many values, making it impossible for Label Encoder to encode them. Using one-hot encoding, categorical data are converted into several binary variables, each of which indicates a possible category value.  4. Separate the data into sets for testing and training: The data was split between training and testing sets using the train\_test\_split function from sklearn.model\_selection. This was carried out in order to train the model on a subset of the data and evaluate its performance on another subset.  No further preparation was done because the provided data was relatively clean and didn't need a lot of preprocessing. |
| **2.b. Feature Engineering** | There was no attempt to generate features in this code. Since the attributes were present in the dataset, the code simply eliminated the columns ID, age\_band, gender, car\_model, and car\_segment that weren't deemed required for the study.  Nothing in the code was removed from functionality. Thus, it's possible that some characteristics were removed before the original dataset's analysis was done. |
| **2.c. Modelling** | For this experiment, I trained a Logistic Regression model. I selected this model because it is straightforward to interpret and performs effectively with binary classification problems. Additionally, because Logistic Regression can handle both categorical and numerical data, it is well-suited for our dataset.  I performed a grid search to optimize several hyperparameters to maximize the accuracy of the Logistic Regression model. The hyperparameters adjustments were:  C: Regularization strength which must be a positive float. Smaller values specify stronger regularization. I tested values of 0.01, 0.1, 1 **and** 10 to see how different levels of regularization affect the model's performance.  solver: Algorithm to use in the optimization problem. I experimented with liblinear and saga, both of which are good choices for small to medium datasets.  penalty: The norm used in the penalization. I tested both l1 (lasso) and l2 (ridge) penalties to explore their impact on the model, especially in terms of feature selection and model simplicity.  The grid search yielded the following optimal hyperparameters: {C=1}, {solver='liblinear'}, and {penalty='l1'}. With these settings, the Logistic Regression model achieved an accuracy score of 89.7% on the test set. This configuration provided a good balance between model complexity and generalization, preventing overfitting while maintaining a high level of predictive performance.  I didn't train any further models for this experiment because the decision tree model performed well on the dataset and was sufficient for our needs.  One hyperparameter that may be important for more investigation is {max\_features}, which sets the maximum number of features to consider when finding the optimal split. Overfitting can be reduced and generalization performance can be improved as a result. |

|  |  |
| --- | --- |
| 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** | Regarding the technical performance of our Logistic Regression model, the final scores were highly revealing. The model achieved an impressive test accuracy of approximately 97.8%, with precision at about 83.4%. However, it's noteworthy that the recall score was significantly lower, at only 20%. This suggests that while the model is highly accurate overall, it has a substantial limitation in its ability to identify all relevant cases of the positive class.  The F1-score, which balances precision and recall, was approximately 32.2%, indicating that the low recall significantly impacted the model's overall performance. This is corroborated by the ROC AUC score, which was at 90.3%, reflecting the model's good performance in distinguishing between the classes but also hinting at the potential for improvement, particularly in recall.    **Confusion Matrix for Customer Repurchase Predictions**   |  |  |  | | --- | --- | --- | |  | Predicted No | Predicted Yes | | Actual No | 38331 | 41 | | Actual Yes | 824 | 206 |   The number of customers who were correctly predicted to repurchase (true positives) is 206. The number of actual positives (all customers who would repurchase) is the sum of true positives and false negatives, which is 206 (true positives) + 824 (false negatives) = 1030.  Therefore, out of the total number of customers who actually would repurchase, the model has correctly identified 206. |
| **3.b. Business Impact** | With a 97.8% accuracy, the logistic regression model is a robust tool for targeting potential car buyers. However, the model’s occasional errors, false positives and negatives could lead to missed sales or wasted marketing spend. Therefore, the business should judiciously employ the model, focusing on likely buyers with limited budgets or expanding the target audience when more funds are available, always mindful of the model's predictive limitations. |
| **3.c. Encountered Issues** | We encountered some difficulties while using logistic regression to navigate the experiment. A unbalanced dataset that heavily favors one class may have skewed the results. We need more clever strategies to level the playing field, such as oversampling or creating entirely new, fake data points. The issue of missing values is another. While it may have been easier to just eliminate them, using statistics such as the mean to fill in the blanks might help maintain the focus of our data.  Plus, we evaluated our model only based on accuracy. However, being correct isn't enough; what matters is how you get the answer properly. Metrics that display the entire picture include F1 scores and ROC-AUC. Finally, to really refine our model, we need to go further into the reasons behind some of the failed predictions. We'll focus more on these areas going future to create a more sophisticated prediction machine. |

|  |  |
| --- | --- |
| 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.  1.With a high accuracy score of 98.81%, the Decision Tree model demonstrated its ability to accurately forecast which customers are most likely to make repeat purchases  2. In fine-tuning the logistic regression model, we focused on honing key hyperparameters to optimize performance. We adjusted the regularization strength (C), the solver algorithm, and the penalty type. The GridSearchCV identified the best performers: C was set at 1 to balance bias and variance, 'liblinear' was chosen as the solver for its efficiency with smaller datasets, and the 'l1' penalty was used to encourage sparsity in the model, effectively performing feature selection. This combination helped us strike a balance between model complexity and predictive power, crucial for achieving robust performance on our binary classification task. |
| **4.b. Suggestions / Recommendations** | 1. Refinement of Feature Set: Enhance predictive accuracy by expanding and refining the model's features to uncover deeper customer insights. 2. Robust Hyperparameter Tuning: Focus on fine-tuning feature selection techniques and regularization to improve the model's generalization and reduce overfitting. 3. Incorporate Time-Series Analysis: Integrate time-series analysis to capture temporal trends, potentially improving forecasting accuracy for dynamic data sets. 4. Analysis of Model Misclassifications: Predicted customers who did not repurchase and predicted non-repurchase customers who did were the primary underperforming situations. To enhance the model's performance and comprehend the underlying reasons of these inaccurate forecasts, more research is required.   Taking everything into account, the trial was successful in achieving its commercial objective of predicting the repurchasing behaviour of customers. Further study is required to improve the functionality of the model and gain a deeper comprehension of the factors impacting customers repurchase behavior. Subsequent studies might examine different models (such Gradient Boosting or Random Forest) and try out different feature engineering techniques. |