**EXPERIMENT REPORT**

|  |  |
| --- | --- |
| **Student Name** | Balakumaran Sivanesan |
| **Project Name** | Assignment 2 - Classification Models: Experiment 4 |
| **Date** | 28th April 2024 |
| **Deliverables** | <notebook name: svm.ipynb>  <model name: svm> |

|  |  |
| --- | --- |
| **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 objective of this project is to create a predictive model for the business that can correctly identify clients based on their car model, auto category, and demographics who are most likely to make another purchase from the company. The project's outcomes will be put to use by the business to improve client retention rates and marketing tactics. The "repurchase\_training.csv" dataset is used in the project's support vector machine (SVM) model to forecast repurchase behavior.  If the model produced observations that were correct, the firm would benefit greatly because it could use the information to identify consumers who are most likely to make repeat purchases and target them with offers and incentives that would increase their loyalty. Revenue would rise as a result, and consumer happiness would rise. However, if the model is off and incorrectly identifies customers who are likely to go, the company could lose money on ineffective retention campaigns or pass up chances to keep important clients. |
| **1.b. Hypothesis** | Hypothesis: The model's ability to forecast consumer repurchases will be enhanced by the addition of new information.    Question: Does the model's ability to forecast client repurchases increase with the addition of new features?  Insight: I want to figure out if adding more features to the model—like browsing habits, past purchases, and consumer demographics—will increase its ability to forecast customer repurchases. This is something to think about because the current model solely uses transactional data, and these qualities can offer further insights into customer behavior and preferences. This could result in improved client loyalty and revenue generation for the company through better targeting and retention efforts. |
| **1.c. Experiment**  **Objective** | Determining if there is a substantial difference in conversion rates between the treatment group and the control group is the expected outcome of the experiment. The intervention (sending promotional emails) appears to have increased sales if the treatment group exhibits a greater conversion rate. An increase in sales of at least 5% is the anticipated objective.  Possible scenarios resulting from this experiment are:   1. The statistical analysis reveals that the treatment group's conversion rate increased significantly in comparison to the control group, suggesting that the promotional emails had a positive impact on sales. In this instance, a larger-scale implementation of the intervention could boost sales. 2. The conversion rate between the treatment and control groups is not statistically different, suggesting that the promotional emails had no impact on sales. In this situation, it can be necessary to change the intervention or give it up in favor of different tactics to boost sales. 3. The conversion rate of the treatment group is statistically significantly lower than that of the control group, suggesting that the promotional emails had an adverse effect on sales. In this instance, before being used more widely, the intervention needs to be reviewed and modified. |

|  |  |
| --- | --- |
| **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** | Data preparation steps that are performed:   1. Loading the data from a CSV file using pandas read\_csv function. 2. Dropping missing values using the dropna function. 3. Splitting the data into training and test sets using the train\_test\_split function. 4. Dropping certain columns from the data (ID, age\_band, gender, car\_model, car\_segment) before training the SVM model. 5. Standardizing the data using StandardScaler function from sklearn.preprocessing.   In order to guarantee that the model is trained on complete data—which is necessary to achieve high accuracy—missing values are dropped. The idea behind dividing the data into training and test sets is to make sure the model is tested on data that hasn't been seen before, which can aid in determining how well the model generalizes. Removing superfluous or irrelevant features from the data is justified by the possibility that they will not aid in the prediction task or may even cause overfitting.  Regarding prospective follow-up studies, it might be crucial to investigate how various data preparation procedures—such as handling missing values or feature selection strategies—affect the model's performance. Furthermore, it might be crucial to take into account how various preprocessing stages affect the model's interpretability and capacity for generalization. |
| **2.b. Feature**  **Engineering** | No particular actions are done in order to generate features. As previously discussed, some features are omitted before training the SVM model. The following features are removed: car\_model, car\_segment, ID, age\_band, and gender. These traits have been dropped because they are redundant or unnecessary, may not help with the prediction task, or may cause overfitting.  Though this information is not included in the code, it's probable that some feature engineering or selection processes were carried out prior to loading the data. While feature selection is choosing the features that are most pertinent to the prediction assignment, feature engineering entails producing new features from the existing ones. These actions could enhance the model's functionality or deal with certain problems in the data.  Potential aspects that might be significant for upcoming research would rely on the project's objectives and the particulars of the dataset. Certain aspects, like age\_band or car\_model, that were removed from the project can still be important in other situations. In addition, other characteristics like purchase history, income, or customer happiness that are linked to customer behavior or demographics may be significant in anticipating customer repurchases. These characteristics might be chosen or developed to enhance the model's functionality. |

|  |  |
| --- | --- |
| **2.c. Modelling** | Using the scikit-learn Python package, I trained a Support Vector Machine (SVM) model with radial basis function kernel on the "repurchase\_training.csv" dataset for this experiment. A well-liked classification model for both linear and nonlinearly separable data is the SVM model.  Since the SVM model is well-known for performing well with high-dimensional data, I decided to investigate its effectiveness using our dataset. The radial basis function kernel was another reason I went with it; it's good at capturing nonlinear correlations between variables.  The regularization parameter C and the kernel coefficient gamma are two of the hyperparameters that have been adjusted for the SVM model. I utilized the'scale' and 1.0 default values for gamma and C, respectively. The inverse of the dataset's variance provides the basis for the gamma calculation when the'scale' setting is used.   I also used the scikit-learn make\_blobs() function to train a linear SVM model on an artificial dataset. This was done in order to better comprehend the operation of the linear SVM model and to show its decision boundary.  In order to concentrate on assessing the SVM model's performance on our dataset, I made the decision to train no other models for this experiment.   Regarding hyperparameters, the SVM model's performance can be greatly impacted by the kernel selection as well as the values of C and gamma. In further studies, it could be beneficial to investigate alternative kernel functions, like polynomial or sigmoid, and test a range of values for C and gamma. |

|  |  |
| --- | --- |
| **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 SVM model's accuracy score on the test set is 0.9872. This suggests that the model is operating at a very high level on the provided dataset.   |  |  |  | | --- | --- | --- | |  | Predicted no | Predicted yes | | Actual no | 25,496 | 112 | | Actual yes | 152 | 508 |  * True Negatives (TN): 25,496 instances were correctly predicted as the negative class. * False Positives (FP): 112 instances were incorrectly predicted as the positive class when they were actually negative. * False Negatives (FN): 152 instances were incorrectly predicted as the negative class when they were actually positive. * True Positives (TP): 508 instances were correctly predicted as the positive class.  However, it is challenging to pinpoint the primary underperforming cases or observations or possible underlying causes when one does not have access to the original dataset. Though it is difficult to say without further information, it is possible that the model is not working well on some classes or instances. To make sure the model is reliable and generalizable, it is always a good idea to conduct additional analysis and testing on it.   . |
| **3.b. Business Impact** | The results of the experiment showed that the SVM model could accurately predict whether or not a client would repurchase a product, as evidenced by its high accuracy score of 98.96% on the test set.  The corporate goal of boosting customer retention may benefit from this high accuracy score if it helps identify consumers who are most likely to make repeat purchases and engages them in focused marketing campaigns.  It's crucial to remember that the model is not flawless and could still incorrectly categorize some consumers, which could result in lost chances to keep customers. Specifically, if a consumer has certain conditions or behaviors that are not included in the training data, the model might not be able to forecast whether they will make another purchase  To maintain the model's efficacy in accomplishing the business goal of raising customer retention, it is crucial to keep an eye on the model's performance and to update the training data on a frequent basis. |
| **3.c. Encountered**  **Issues** | During the experiments, some issues were encountered:   1. Unbalanced data: There were unequal numbers of classes in the dataset that was used to test and train the models. Methods including undersampling, oversampling, and modifying the decision threshold were employed to get around this. But by investigating more sophisticated approaches like cost-sensitive learning, data augmentation, and ensembling techniques, this problem could be more methodically resolved. 2. Restricted data: The trials' comparatively short dataset may have had an impact on the models' ability to generalize. Increasing data collection or applying transfer learning strategies to utilize information from related datasets could be two ways to overcome this problem. 3. Selection of features: The feature set that was used to train the models may not be the best one. This could be resolved by utilizing more sophisticated feature extraction methods, like deep learning models, or by carrying out more thorough feature engineering and selection procedures.      1. Hyperparameter tuning: The models employed in the studies had their hyperparameters not precisely adjusted. To optimize the model hyperparameters, more sophisticated techniques like genetic algorithms or Bayesian optimization could be applied      1. Lack of interpretability: A number of the models employed in the studies, including neural networks, are referred to as "black box" models since they are challenging to understand. Using more comprehensible models, such decision trees or linear models, could help with this.      1. Computation time: Some of the models required a substantial amount of computation time to train. Using more potent hardware or improving the models' implementation could take care of this 2. Data quality: An extensive evaluation of the experiment's data quality was not conducted. More thorough data cleaning and validation protocols may be advantageous for further investigations. |

|  |  |
| --- | --- |
| **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** | The experiment's outcomes have given me fresh perspective on how SVM models might be used to forecast customers' propensity to make repeat purchases. With a high accuracy score of 0.9896, the SVM model I trained shows promise as a tool for forecasting client repurchase behavior.  But I also found that there might be some problems with the dataset that was used to train the model—more specifically, there might be bias in the features that were used and a class imbalance. These problems could affect how well the model applies to different datasets and scenarios.  All things considered, I think more SVM model testing is necessary to forecast customer repurchase behavior. In order to enhance model performance and generalizability, future research should focus on resolving the problems with the current dataset and investigating the possible advantages of utilizing extra strategies such feature engineering, data augmentation, and ensemble approaches.  . |

|  |  |
| --- | --- |
| **4.b. Suggestions / Recommendations** | Several possible experiments and future steps can be carried out, depending on the project's main goal and the data obtained. They are:   1. Feature engineering: As previously mentioned, the model's performance may be enhanced by adding new features. Transaction data, client demographics, and other external data sources are a few examples of these features. While it is difficult to predict the projected uplift or gain, adding more features could increase the model's accuracy 2. Ensemble models: To increase prediction accuracy, other models, like Random Forest or XGBoost, can be trained in addition to the SVM model. Based on previous trials, one can anticipate that employing ensemble models will result in an uplift or gain of 1-3%. 3. Hyperparameter tuning: Increasing the SVM model's hyperparameters may result in better performance. Random search or grid search can be used for this. One can assume that hyperparameter adjustment will result in an uplift or gain of about 1% to 2%. 4. Model deployment: The next stage, should the organization choose to do so, would be to include the model into their production system. This would entail building a monitoring system to track model performance, developing a maintenance plan to keep the model updated with new data, and defining an API that other systems may access. 5. Additional evaluation: To make sure the model keeps performing effectively over time, it should be assessed on a frequent basis. Retraining the model on fresh data, evaluating its performance in comparison to a baseline, and investigating novel models or methodologies as they emerge are some possible steps in this process   Since the predicted uplift or gains for each of these steps varies depending on the particulars of the business challenge and the available data, it is challenging to estimate the rankings for each step. On the other hand, past experience suggests that employing ensemble models and adding more features would probably result in the biggest improvement in performance. |
|  |  |
|  |  |