Assignment 1  
Classification Models

short line

Ronik Jayakumar

Student ID: 24680264  
26/04/2024

36106 - Machine Learning Algorithms and Applications

Master of Data Science and Innovation

University of Technology of Sydney

Table of Contents

[**1. Business Understanding 2**](#_if9iwijz2pg3)

[a. Business Use Cases 2](#_bnjfuh5ewogw)

[**2. Data Understanding 3**](#_yg03y3w5hig9)

[**3. Data Preparation 4**](#_tj77rrmixzhi)

[**4. Modeling 5**](#_6vwh14nc44ga)

[**5. Evaluation 6**](#_mngiyx5tet6w)

[a. Results and Analysis 6](#_25v57zsj5m7g)

[b. Business Impact and Benefits 6](#_hkob76wu4d6q)

[c. Data Privacy and Ethical Concerns 6](#_uf1z6gbsejg6)

[**6. Conclusion 8**](#_9hr11g79asdx)

[**7. References 9**](#_ma3y8ytjvp1y)

# Business Understanding

## Business Use Cases

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. Speific objectives of the business are as follows

* Customer identification so marketing efforts and resources can be efficiently used.
* Assist in developing targeted marketing and loyalty programs to retain existing customers.
* Understanding customer repurchase behavior to estimate future sales quantities, allowing for more effective resource allocation and production planning.
* Customer segmentation based on repurchase likelihood allows for personalized communication methods and product offerings for each category, hence boosting overall marketing effectiveness.

Challenges and Opportunities:

* Automotive datasets can contain huge amounts of complicated, diverse data, which pose hurdles for typical analysis approaches.
* Customers make high investments in vehicles, thus accurate forecasts of repurchase behavior are critical to the profitability and long-term success of automakers.
* Market diversity and relatively low turnover rates required the use of data driven decision making.
* Machine learning algorithms excel at understanding complex patterns and relationships with little to no bias in a well-made model. This is an avid requirement in the automotive field.

1. Key Objectives

The key objectives for this experiment are as follows:

* Predictive Analysis: Using previous data, create reliable models to anticipate the possibility that customers would repurchase autos.
* Customer Segmentation: Divide customers into categories based on their purchasing habits and demographics.
* Optimized Marketing tactics: Determine the primary elements driving repurchase behavior to create targeted marketing tactics and boost client retention.
* Business Growth: Improve overall business performance by improving client repurchase rates and revenue through more effective marketing initiatives.

Stakeholders and their requirement analysis has been conducted using the power interest grid.

* Internal office departments that are affected by the project. They are as followd:
  + Sales
  + Marketing
  + Management
  + Customer Service
  + Data Science
* External Stakeholders:
  + Customers
  + Automotive companies
  + Financial institutions

The project addresses requirements in the following ways

* Machine Learning Models: Use machine learning algorithms to analyze previous customer data and create forecast models of repurchase behavior.
* Feature Engineering: Extract important information from the dataset, such as purchase history, demographics, and previous interactions, to improve model accuracy.
* Model Evaluation: To ensure dependability and accuracy in forecasting consumer repurchase behavior, models should be trained and evaluated using rigors validation procedures
* Actionable insights: Transform model predictions into useful information for stakeholders, allowing for focused marketing efforts, personalized customer experiences, and strategic decision-making.

short dash

# Data Understanding

Data Source:

* The dataset consists of 131337 rows and 17 columns.
* The data was provided by the University of Technology Sydney to conduct experimentation on the repurchase data.
* Additional data may have been collected through surveys, customer feedback, or third-party sources.

Data Limitations:

* The data has missing rows in the age\_band and the gender columns.
* These have been dealt with by dropping the age\_band column due to over 90% of the rows being null and replacing the missing gender variables using MICE imputation.
* The categorical features are isolated to find ways in preprocessing the columns.

Features present in the dataset are as follows:

* ID: Unique ID of the customer
* target: Model target. 1 if the customer has purchased more than 1 vehicle, 0 if they have only purchased 1.
* age\_band: Age banded into categories
* gender: Male, Female or Missing
* car\_model: The model of vehicle, 18 models in total
* car\_segment: The type of vehicle
* age\_of\_vehicle\_years: Age of their last vehicle, in deciles
* sched\_serv\_warr: Number of scheduled services (e.g. regular check-ups) used under warranty, in deciles
* non\_sched\_serv\_warr: Number of non-scheduled services (e.g. something broke out of the service cycle) used under warranty, in deciles
* sched\_serv\_paid: Amount paid for scheduled services, in deciles
* non\_sched\_serv\_paid: Amount paid for non scheduled services, in deciles
* total\_paid\_services: Amount paid in total for services, in deciles
* total\_services: Total number of services, in deciles
* mth\_since\_last\_serv: The number of months since the last service, in deciles
* annualised\_mileage: Annualized vehicle mileage, in deciles
* num\_dealers\_visited: Number of different dealers visited for servicing, in deciles
* num\_serv\_dealer\_purchased: Number of services had at the same dealer where the vehicle was purchased, in deciles

Exploratory data analysis:

* EDA is conducted using summary statistics, data types checks, null values checks, and correlation analysis.

short dash

# Data Preparation

The steps taken to prepare the data are as follows:

* Null values handling
  + Age\_band column has over 90% of the rows null. Hence this column has been dropped.
  + The gender column has approximately 50% of its rows as null.
* Categorical Feature handling:
  + Label encoder has been used to encode the categorical features into integer values.
  + The ‘Other’ values in the car\_segment column have been dropped to decrease dataset complexity for model training.
* Target variables
  + The target variables have been explored where a high-class imbalance has been noticed. This has been dealt with in different ways in each experiment to find an effective method.
* Feature Engineering:
  + Feature engineering has been performed on the gender column by using MICE imputation. This predicts and replaces all the null values.
* Correlation analysis:
  + The correlation between the features and the target variable has been explored by using correlation analysis with a heatmap and the ANOVA-F test.
* Hyperparameter tuning:
  + Each experiment has hyperparameter tuning performed depending on the model that has been trained.

Following the above data preprocessing methods, 8 columns have been selected for model training. They are:

* age\_of\_vehicle\_years
* sched\_serv\_warr
* sched\_serv\_paid
* total\_paid\_services
* total\_services
* mth\_since\_last\_serv
* Gender
* Car\_segment

short dash

# Modeling

Experiment 1 (Support Vector Machine or SVM):

* SVM is a supervised machine learning algorithm that classifies data by finding an optimal line or hyperplane that maximizes the distance between each class in an N-dimensional space [1]
* Hyperparameter tuning has been performed on the model to find the most optimal weights.

Experiment 2 (Random Forest):

* Random forest is a commonly used machine learning algorithm, trademarked by Leo Breiman and Adele Cutler, that combines the output of multiple decision trees to reach a single result. [2]
* Parameter tuning includes altering the number of trees, maximum depth, and other hyperparameters to improve model performance.

Experiment 3 (Decision Trees):

* A decision tree typically starts with a single node, which branches into possible outcomes. Each of those outcomes leads to additional nodes, which branch off into other possibilities. This gives it a treelike shape. [3]
* Hyperparameter tuning has been performed like tree depth and minimum samples per split to help reduce overfitting and increase model generalization.

Experiment 4: K Nearest Neighbours (KNN)

* The KNN algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. [4]
* Hyperparameter tuning has been performed on the best number of neighbors to apply into our specific use case.

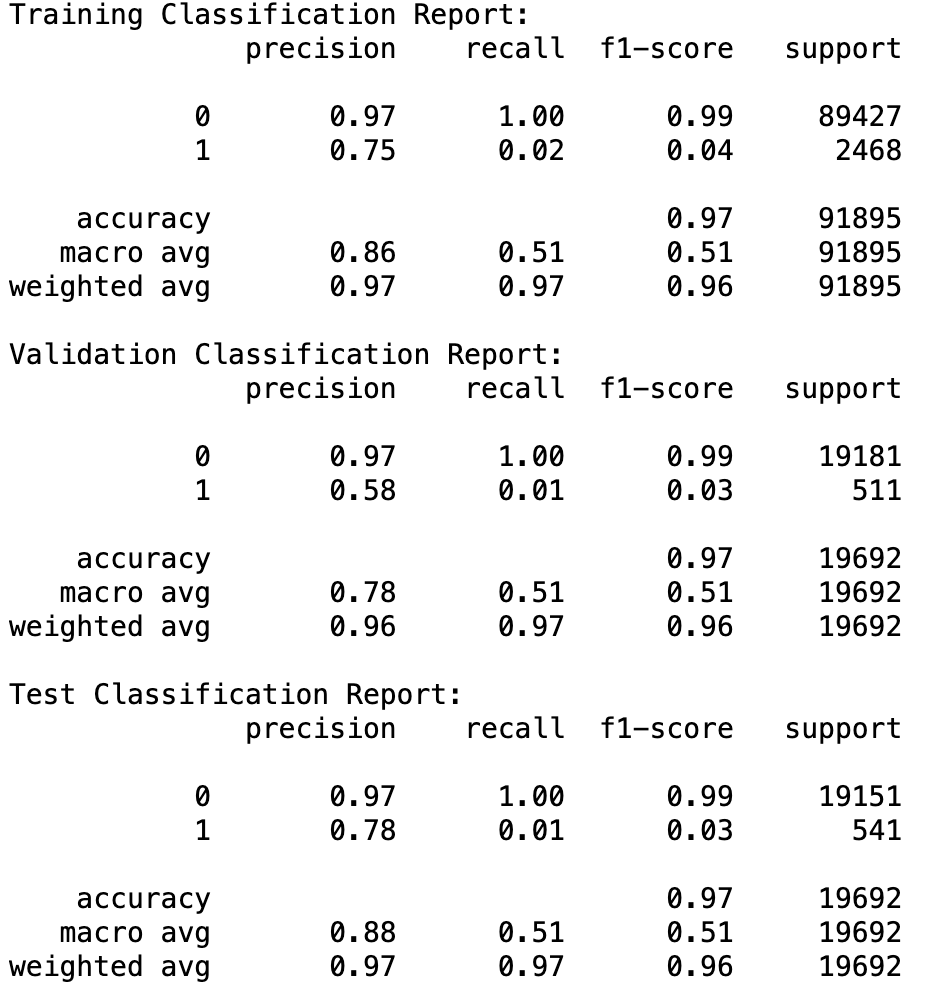
Experiment 5 Gradient Boosting Classifier:

* Machine learning ensemble technique that combines the predictions of multiple weak learners, typically decision trees, sequentially.
* Hyperparameter tuning has been performed.

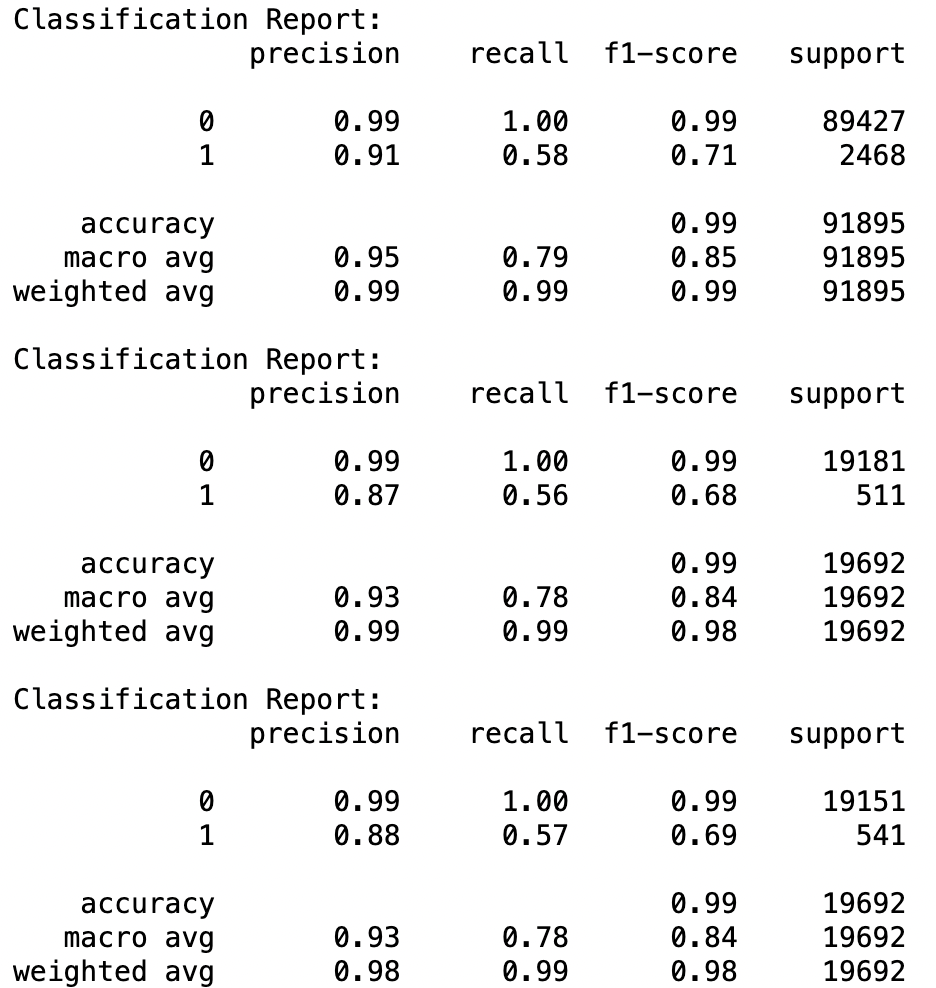
# Evaluation

1. **Experiment 1 (Support Vector Machines):**

Baseline model: Logistic Regression



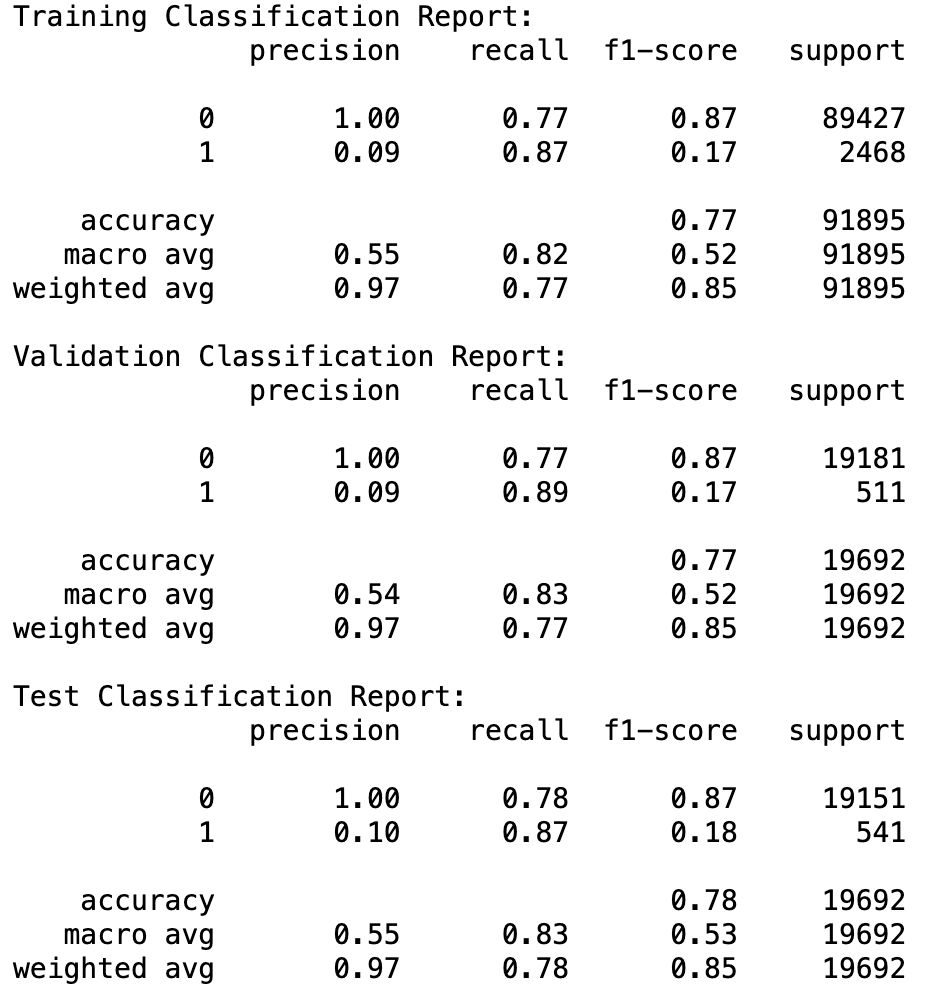
Final model – Support Vector Machines



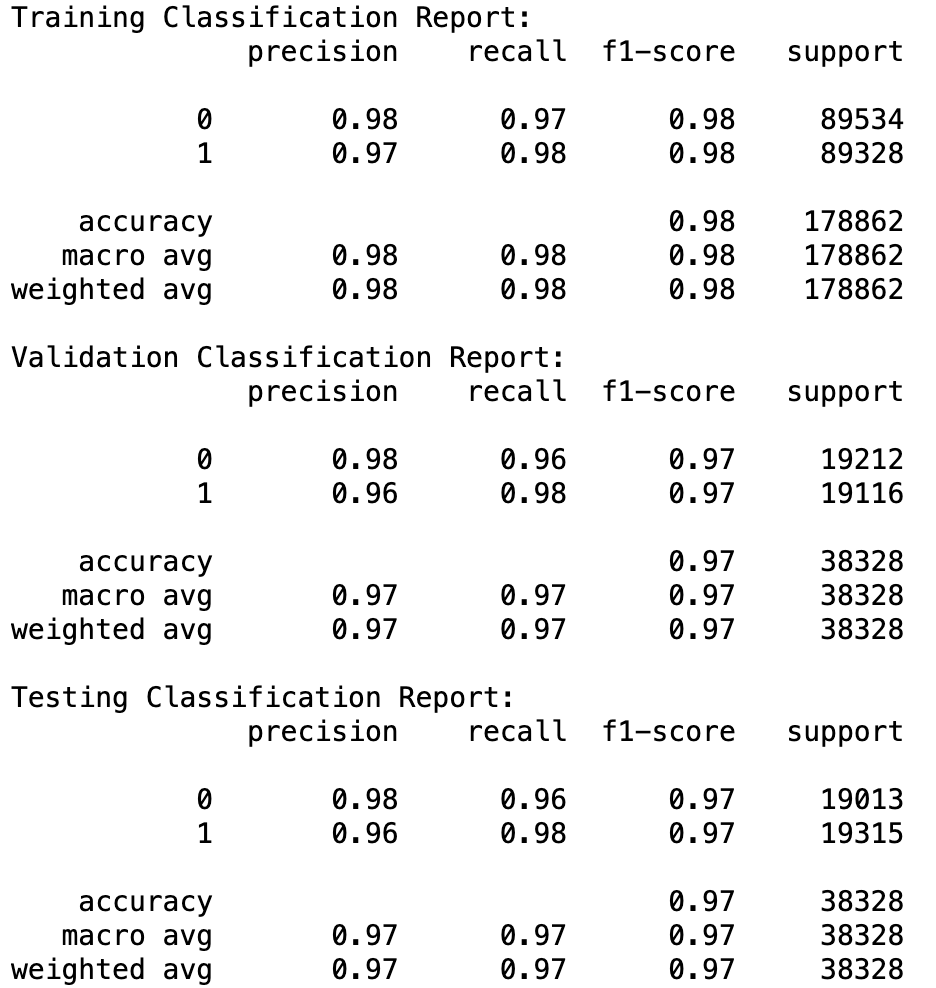
The support vector machines classification report performed well as compared to the baseline logistic regression model. The class imbalance has not been dealt with in this experiment. This shows that deeper models perform considerably well in learning underlying dataset features.

**2. Experiment 2:**

Baseline model – Logistic Regression



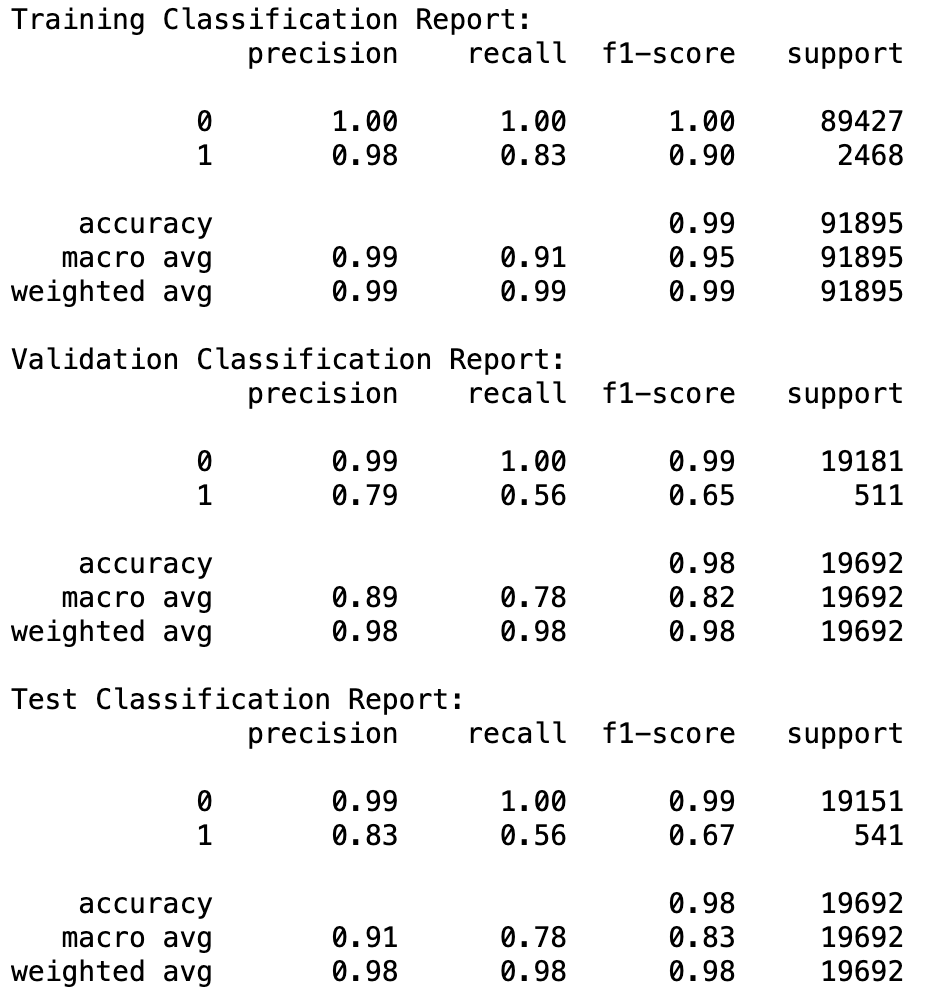
Main model – Decision Trees:



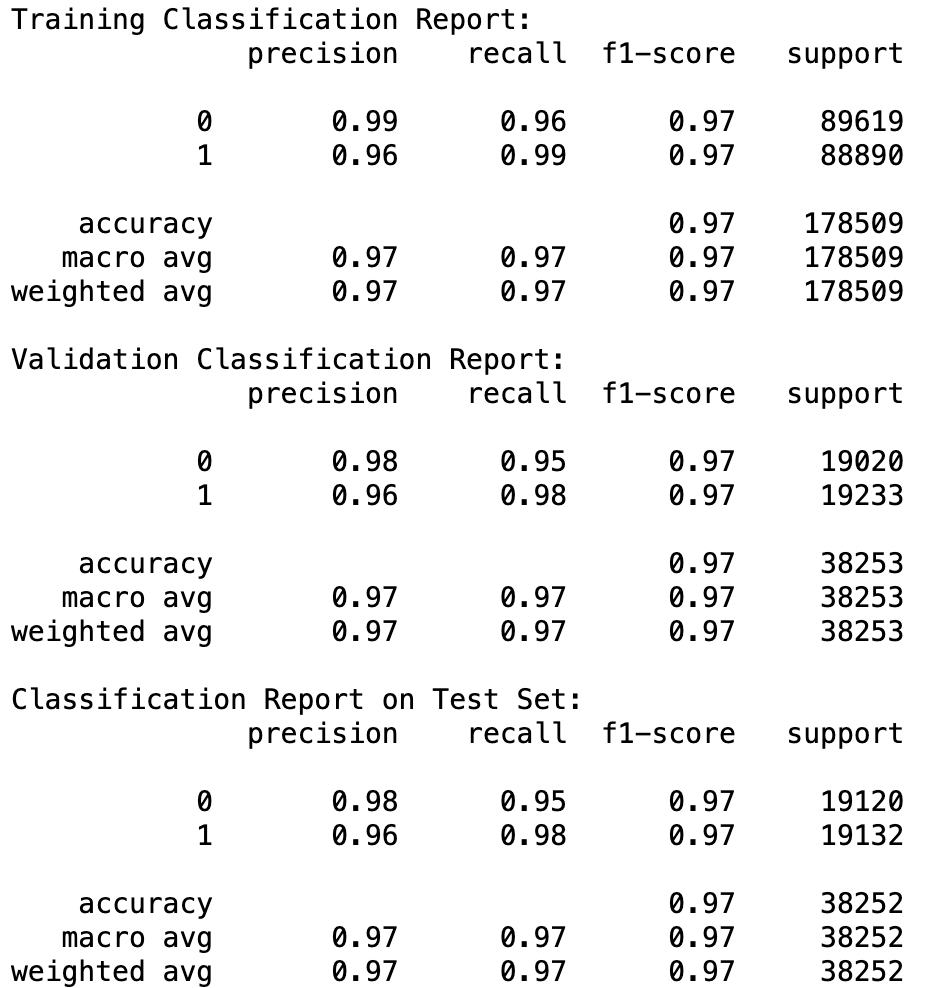
The decision tree classifier performs extremely well especially after class imbalance handling using SMOTE Tomek.

**3. Experiment 3: Random Forest**

Baseline model: Random forest before class imbalance handling



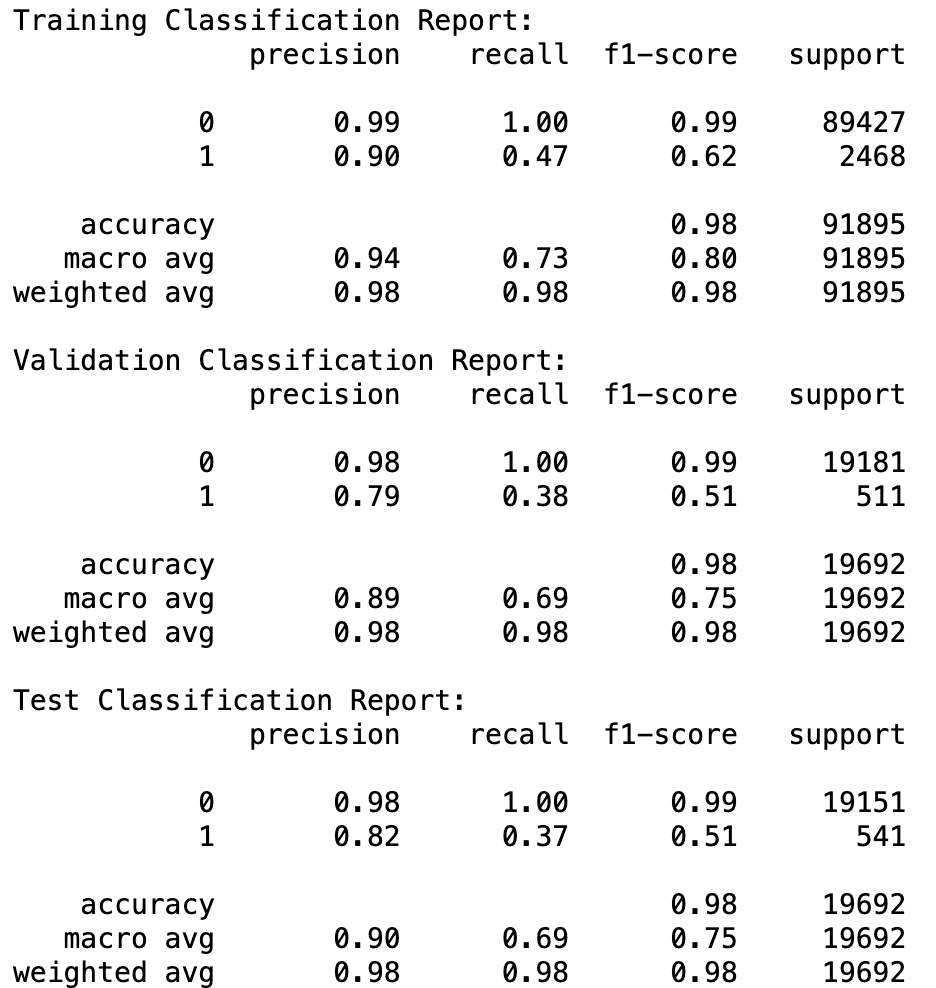
Main model: Random Forest after class imbalance handling



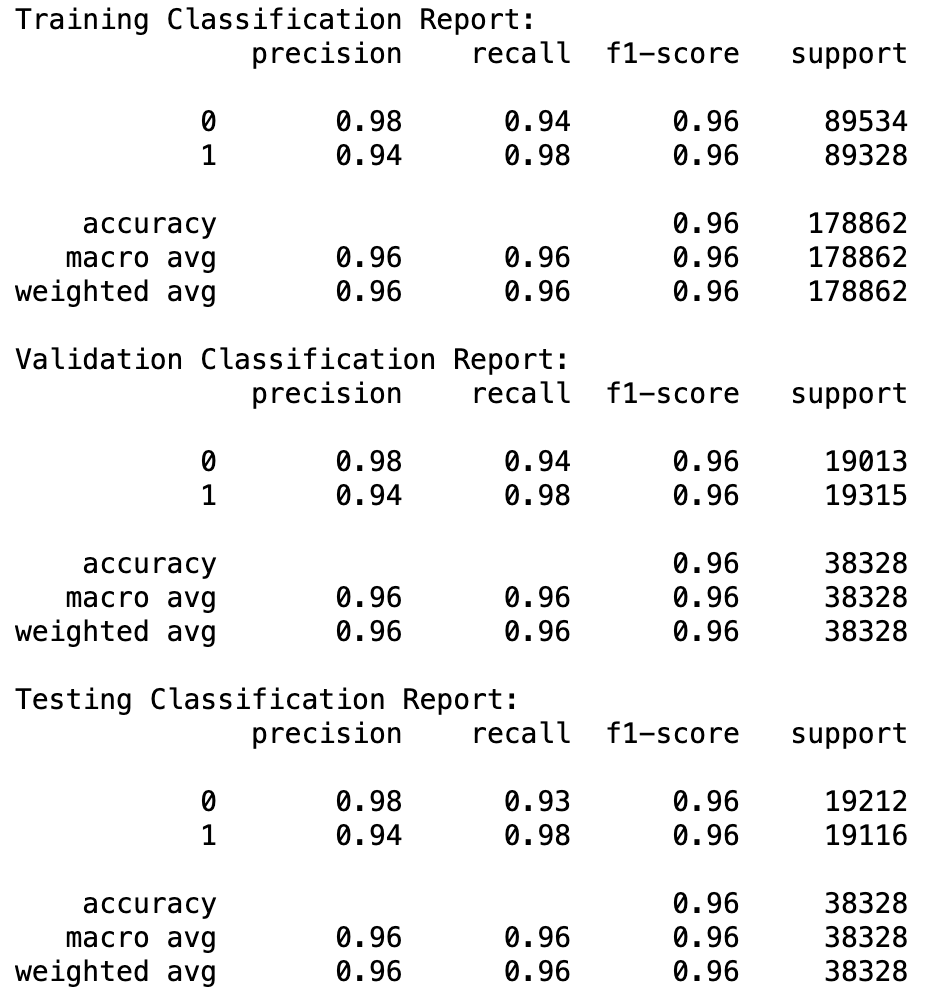
The random forest classifier works extremely well after class imbalance handling. For experiment 3, ADASYN has been for class imbalance handling.

Experiment 4 K Nearest Neighbors (KNN)

Baseline model – KNN Before class imbalance handling



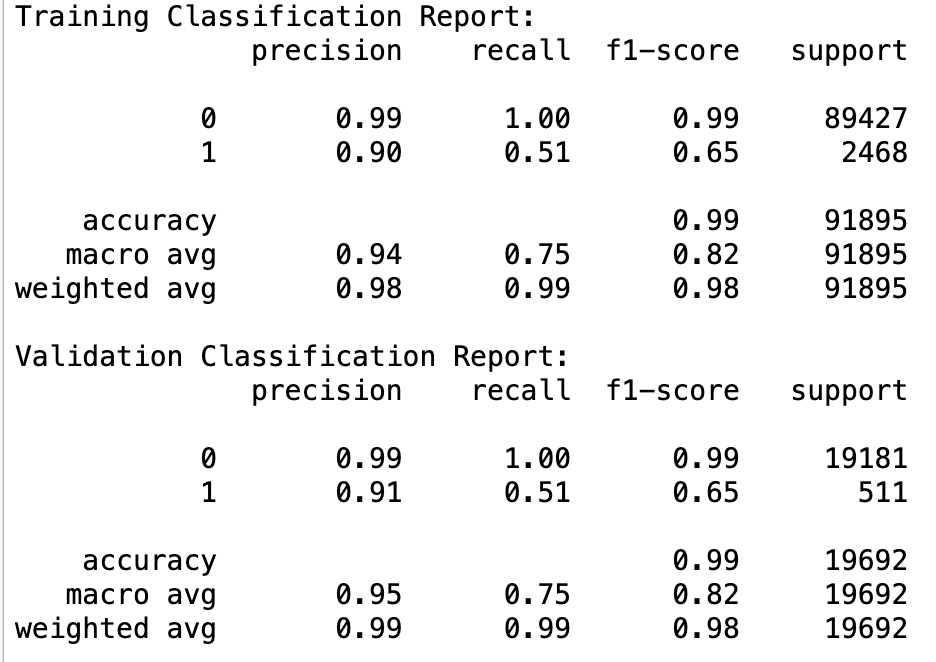
Main model - KNN after class imbalance handling

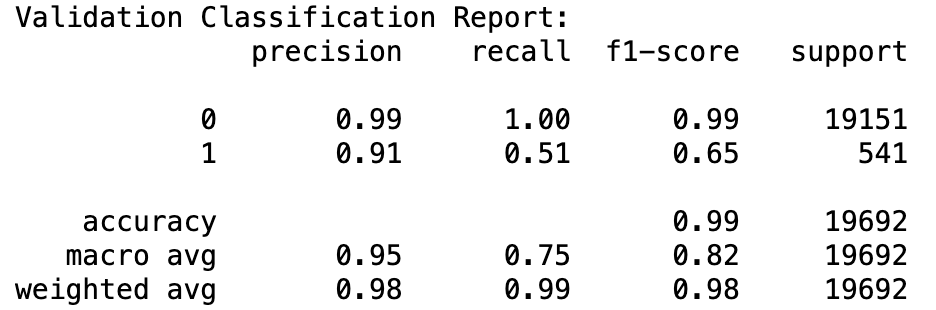


The SMOTE oversampling technique has been used in this experiment. As noticed, the KNN classifier works exceptionally well.

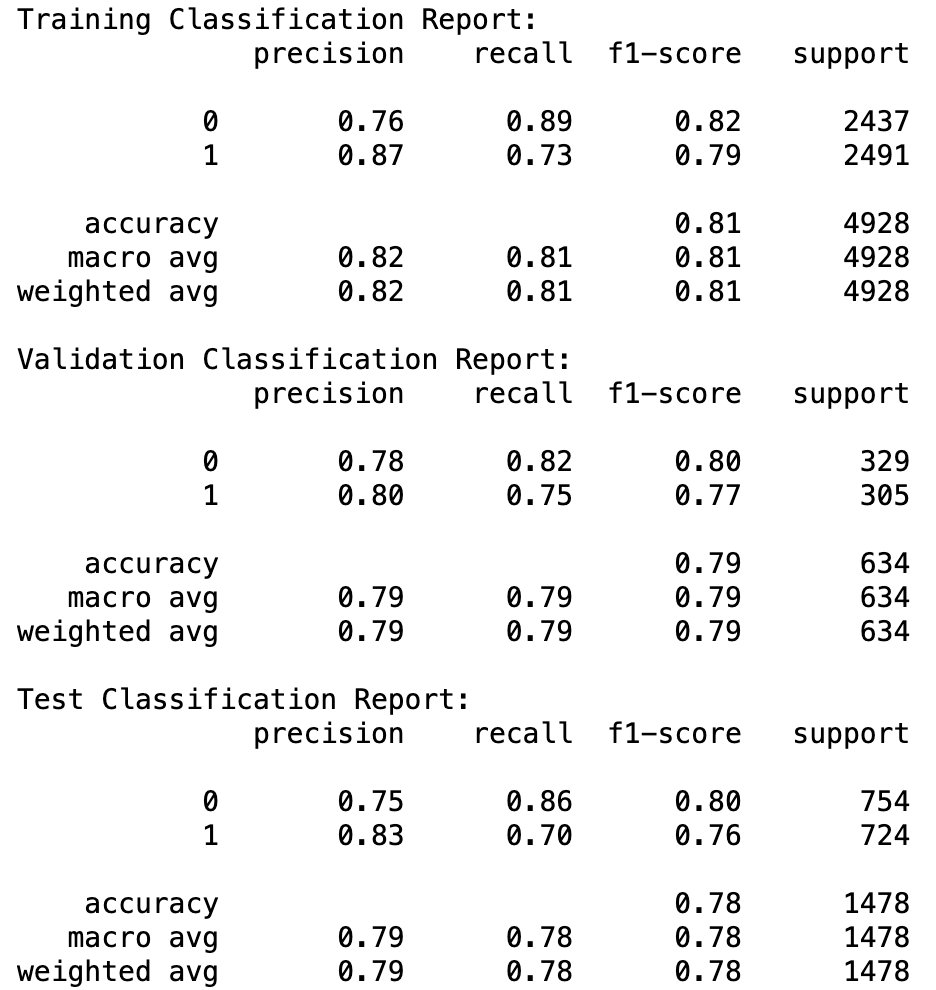
Experiment 5 Gradient Boosting Classifier

Baseline model: Gradient Boosting before class imbalance handling





Main model - Gradient Boosting after class imbalance handling



Experiment 5 uses the near miss under sampling technique. As noticed, the GBC model works well but we have noticed better results with oversampling.

## Business Impact and Benefits

* Informed Decision Making: The model helps to make better judgements by properly estimating the chance of car repurchase, allowing for more targeted resource allocation.
* Enhanced Customer Engagement: Targeted communication and personalised offers to prospective repurchasers boost customer happiness and loyalty.
* Cost Savings: Optimizing marketing expenditure lowers the expenses associated with large campaigns, enhancing ROI
* Accurate forecasts give a competitive advantage, helping to outperform competitors and acquire market share.
* Quantifiable Metrics: Improvements in accuracy, precision, recall, and F1-score show that the model is useful and has an influence on business results.

1. Data Privacy and Ethical Concerns

* Data Privacy: The handling of sensitive consumer information causes privacy and security issues.
* Ethical considerations include ensuring informed consent, eliminating bias in predictions, and protecting against unforeseen effects.
* Privacy Measures: Anonymizing data, maintaining strong security, and adhering to legislation are critical for privacy protection.
* Transparent deployment: Ethical deployment procedures include transparent model deployment, explanations of forecasts, and guaranteeing fairness.
* Impact on Indigenous Communities: Cultural and ethical considerations need to be considered for underrepresented communities such as aboriginals. There needs to be comprehensive studies to ensure all communities, nations, and clans are represented.

short dash

# Conclusion

Key Findings and outcomes:

* The study successfully created machine learning models that predict automobile repurchase behavior.
* The models obtained good accuracy, precision, and recall, demonstrating their efficacy in detecting possible repurchasers.
* Certain characteristics were important in forecasting repurchase behavior and the requirement for continual model optimization.

Reflection on Success:

* The initiative achieved its primary purpose of developing predictive models to aid business choices.
* Stakeholders' expectations for accurate projections were met, resulting in informed decision-making.
* The study highlighted the importance of machine learning in solving business problems and optimizing operations.

Future work and recommendations

* To ensure accuracy and relevance, model performance should be continuously monitored and updated with new data.
* Investigate new characteristics or data sources to improve model predictability and robustness.
* Investigate interpretability strategies to improve model predictions and generate actionable insights.
* Collaborate with domain experts to improve models and customise them for unique company needs.
* Ensure representation of underrepresented aboriginal communities. Ensure studies are conducted in the culturally appropriate manner, respecting indigenous knowledge and values.

short dash

# References

[1] : What are SVMs? - 27th December 2023 - <https://www.ibm.com/topics/support-vector-machine#:~:text=A%20support%20vector%20machine%20(SVM,the%201990s%20by%20Vladimir%20N>.

[2] : What is random forest? - <https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems>.

[3] : What is a Decision Tree Diagram - <https://www.lucidchart.com/pages/decision-tree#:~:text=A%20decision%20tree%20typically%20starts,decision%20nodes%2C%20and%20end%20nodes>.

[4] : What is the k-nearest neighbors (KNN) algorithm? - <https://www.ibm.com/topics/knn#:~:text=The%20k%2Dnearest%20neighbors%20(KNN,of%20an%20individual%20data%20point>.

[5]

short dash