

Assignment 2

Classification Models

Shaqran Bin Saleh
Student ID: 25010238
26-04-2024

36106 - Machine Learning Algorithms and Applications
Master of Data Science and Innovation
University of Technology of Sydney

Table of Contents

Contents

Business Understanding	2
Business Use Cases.....	2
Key Objectives.....	3
Data Understanding.....	5
Data Preparation.....	12
Modeling	14
Evaluation.....	17
a. Results and Analysis.....	17
b. Business Impact and Benefits	21
c. Data Privacy and Ethical Concerns	23
Conclusion	24
References.....	25

Business Understanding

Business Use Cases

Our project was motivated by the integration of machine learning algorithms in making data driven business decisions. Here in case of car repurchasing, the dataset provides a myriad of vehicle characteristics and according to it the customers repurchasing decision. Through the help of this project the following business use can be generated:

1. Customer Retention

Through this project the identification of customers with a high probability to repurchase vehicles allows businesses to tailor campaigns effectively, maximizing customer retention.

Machine learning models used in our project can analyze past purchase data to predict repurchase probabilities among customers.

2. Inventory Optimization

Understanding customer preferences and likely purchase patterns aids in optimizing inventory levels and logistical operations, ensuring dealerships stock vehicles that meet current market trends.


With our project, business can analyze customer preferences and past purchase patterns to forecast which vehicle features and models are likely to be most popular.

3. Strategic Campaigns

The business can efficiently allocate marketing resources to campaigns that will generate the highest return on investment (ROI) is very much important.

As there are use to this project to business. The business might face certain challenges:

- **Data Quality and Integrity:** The effectiveness of machine learning predictions depends on the quality and completeness of the data it will be using. Factors such as missing values, inconsistencies in data entry, and outdated information can significantly hinder model performance.
- **Model Scalability and Maintenance:** Developing models for deployment that not only perform well with historical data but also can handle and adapt to new and evolving datasets presents a significant challenge. Additionally, models require regular updates and maintenance to ensure they remain effective as customer trends and market conditions shift as data is always update in real world scenarios.



The business will be benefitted greatly through proper deployment and integration. The business flourish because the approach can lead to a deeper understanding of customer needs, enabling the development of more effective sales strategies. Also through this approach we can achieve more precise targeting and personalization in marketing campaigns, superior customer service, and more effective inventory management.

Key Objectives

The objectives of this project can be listed as follows:

- Evaluating classification models for repurchase prediction of customers

We will explore and compare the performance of various classification algorithms, such as K-Nearest Neighbors (KNN), Support Vector Machines (SVC), Decision Trees, Random Forest, and Extra Trees.


Our primary goal is to identify the model that delivers the most accurate predictions of customer repurchase behavior. We will evaluate model performance using established metrics like accuracy, precision, and F1-score.

- Optimizing model performance through hyperparameter tuning

Hyperparameter tuning involves configuring specific algorithm settings to optimize prediction accuracy. This process will enable us to enhance the model's reliability and ensure it can be confidently used to predict customer repurchase behavior.

- Identifying key factors in repurchase decisions

Gaining insights into these critical factors will be instrumental in informing targeted marketing strategies and product offerings. This will ultimately allow us to develop initiatives that effectively boost customer retention.



Through the deployment of this project, the following are our targeted stakeholders:

- Marketing Department:

Machine learning models will identify key variables influencing customer repurchase likelihood. These insights will be used to help marketing initiatives for specific customer segments, enhancing campaign effectiveness and return on investment (ROI).

- Inventory Team:

Our project will analyze data to determine which vehicle features are most closely associated with increased repurchase rates. This information will guide the development of vehicle feature that resonate with the target market, potentially leading to higher customer retention.

- Sales Teams:

Classification models will be deployed to predict which customers are more likely to repurchase. Armed with these predictions, sales representatives can proactively engage these individuals, potentially increasing sales and customer loyalty.

- Senior Management:

The data gained from the project will support informed decision-making regarding resource allocation, market positioning, and customer relationship strategies which will help the senior management of the business to take a data driven decision making approach.



Data Understanding

The dataset provided to us for this project was for analyzing customer repurchase likelihood behavior of a vehicle. The dataset can be categorized as CRM(Customer Relationship Management) data. Where we were given various features of different cars and services done by customers and the target column signifying the customer's behavior of repurchasing a car.

There are various limitations:

- The first is after we look into the dataset, we find that we have some missing values which needed to be addressed crucially.
- Imbalance in target variable “target” distribution. The majority class “0” outweighs the minority class “1”. This imbalance can lead to performance metrics to be misleading.

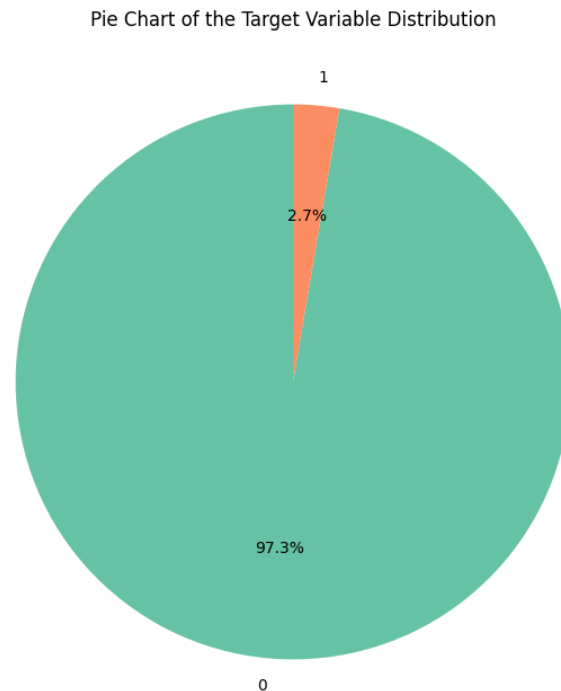


Figure 1: Pie chart showing distribution of Target

The dataset contains the following features:

- age_band: Categorical variable indicating the age group of the vehicle owner. Important for segmenting customer behavior by age.
- gender: Male or Female gender of the customers.
- Vehicle Characteristics:
 - car_model: The model of the car.
 - car_segment: The type of vehicle categorized into types like LCV, Small/Medium, Large/SUV.
 - age_of_vehicle_years: The age of the customer's last vehicle.
- Service Interaction Metrics:
 - sched_serv_warr: Number of scheduled services under warranty. Reflects the initial quality and warranty terms.
 - non_sched_serv_warr: Number of unscheduled services during the warranty period, indicating issues or concerns.
 - sched_serv_paid and non_sched_serv_paid: Services paid for by the customer
 - total_paid_services: Amount paid in total for services by the customer.
 - total_services: Total services availed by the customer.
 - mth_since_last_serv: Time (the number of months) since the last service, a factor in vehicle maintenance and performance.
 - annualised_mileage: Yearly mileage.
- num_dealers_visited: Number of different dealers visited for servicing by the customer.
- num_serv_dealer_purchased: Reflects Number of services had at the same dealer where the vehicle was purchased.

In order to gain better understanding of the dataset, a series of exploratory data analyses were performed.

- Histograms and box plots were employed to visualize the distribution of numerical variables against our target variable “target”, such as total_services, mth_since_last_serv, and annualised_mileage.
- Count plots were constructed to examine the frequency of each category within categorical variables against our target variable “target”. Like gender, age_band, car_segmant and car_model.
- Pie charts were utilized to depict the proportion of observations falling into each category of the target variable. This visualization served to illuminate any potential class imbalance or bias within the dataset.

After performing some analysis on our data, the following insights can be drawn:

- The Gender by Target plot shows that males are the dominant gender in both class of the target and the count target 0 of not repurchasing is greater than ones who repurchased.

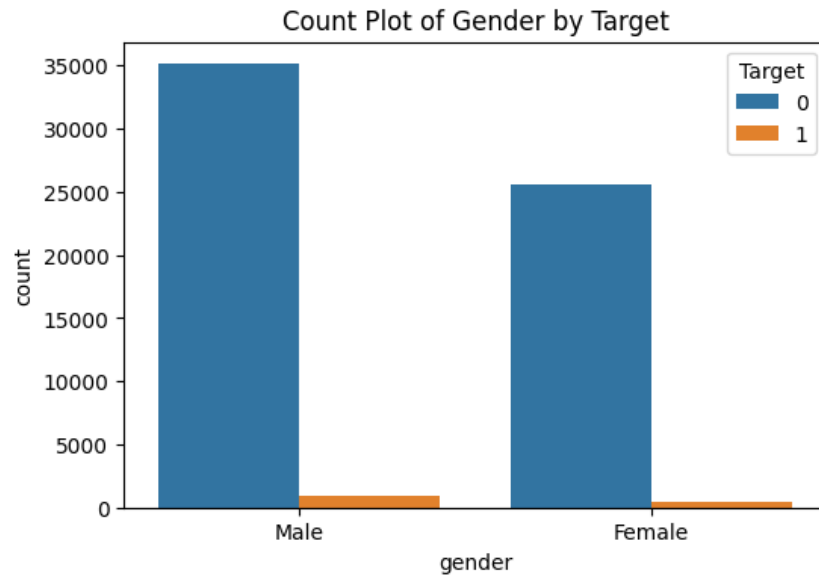


Figure 2: Bar chart showing Gender count by Target

- The Car Type by Target plot shows that the Small/Medium car type is the most numerous in both categories, with Target 0 being substantially higher than Target 1 across all segments

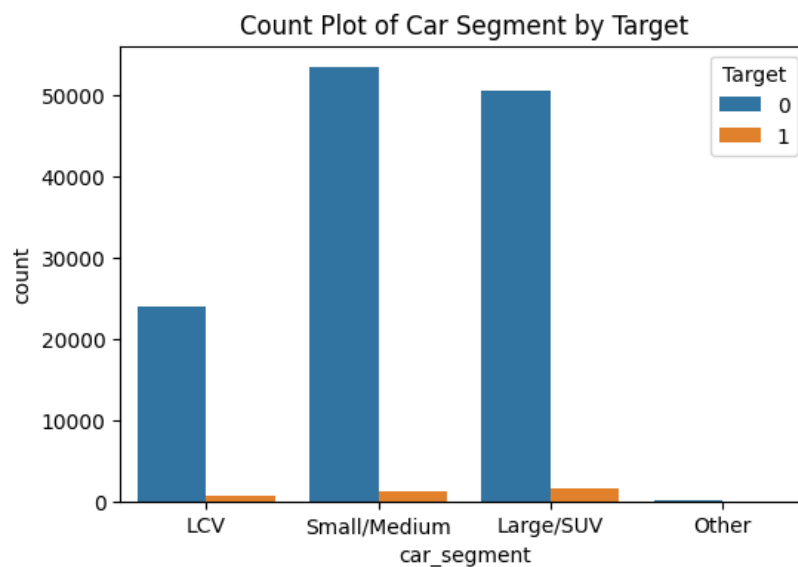


Figure 3: Bar chart showing Car Type count by Target

- The Age of Vehicle by Target box plot shows that target 0 tends to have older vehicles with a narrower interquartile range (IQR), whereas Target 1 has a wider range of vehicle ages, indicating more variability in this category.

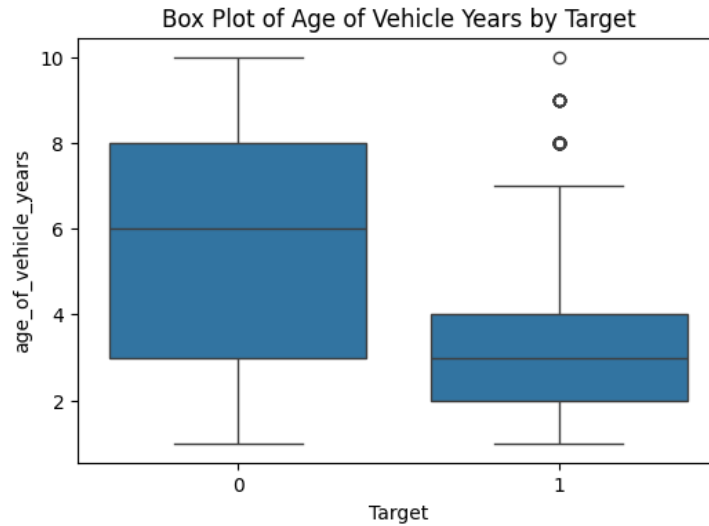


Figure 4: Chart showing Age of customer's last vehicle grouped by Target

- The Total Paid Services by Target chart shows that the number of paid services is fairly similar for both targets, with the median value being higher for Target 0.

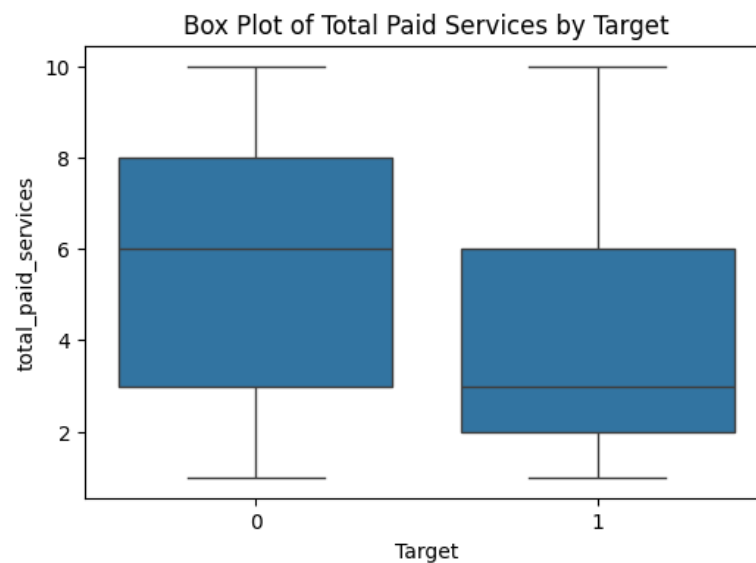


Figure 6: Chart showing Amount paid in total for services grouped by Target

- The distribution of car model shows Model_2 being the most frequently occurring car model. Other models, such as Model_17, Model_18, and Model_19, appear far less frequently.

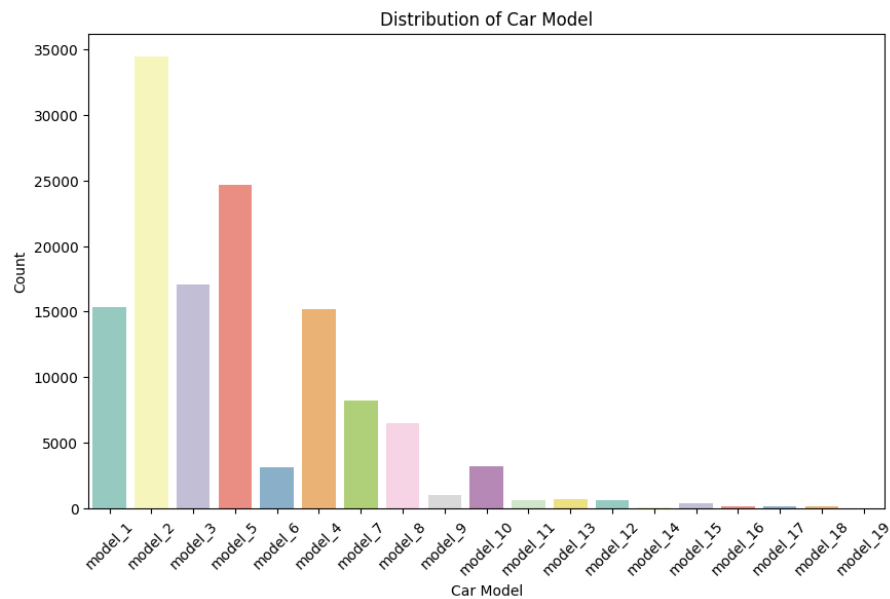


Figure 7: Bar chart showing counts of various car models

- An examination of age bands demonstrates a skew towards middle-aged demographics, with the 45-54 age band being the most prominent. The 75+ age group shows the lowest representation.

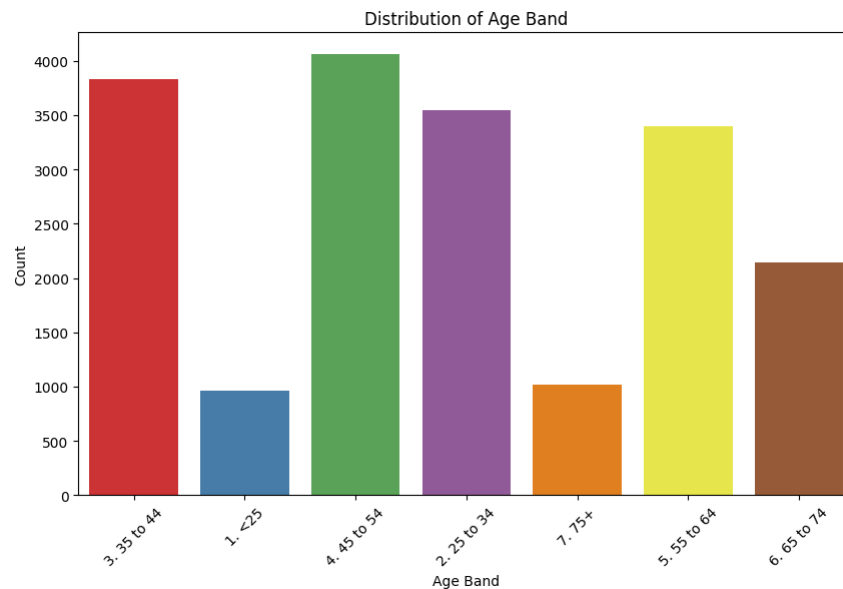


Figure 8: Bar chart showing counts of various age bands

- The pie chart depicting the target variable distribution confirms a significant imbalance within the dataset. Target 0 holds a dominant share of 97.3%, while Target 1 represents only a small fraction (2.7%). This imbalance may require consideration in further analysis.

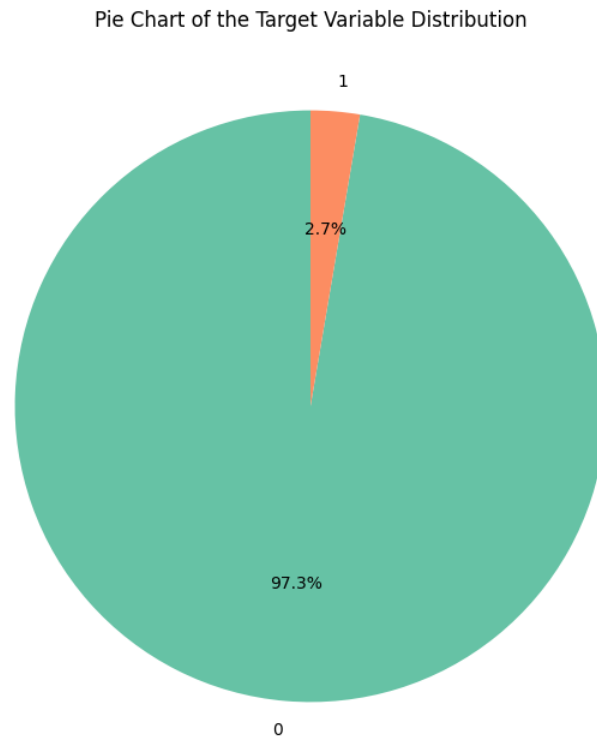


Figure 9: Pie chart showing distribution of Target

- After performing the feature importance method on the training data. We found that the `age_vehicle_years` which is the age of the customer's last vehicle is the leading feature on which the variability of repurchasing depends.

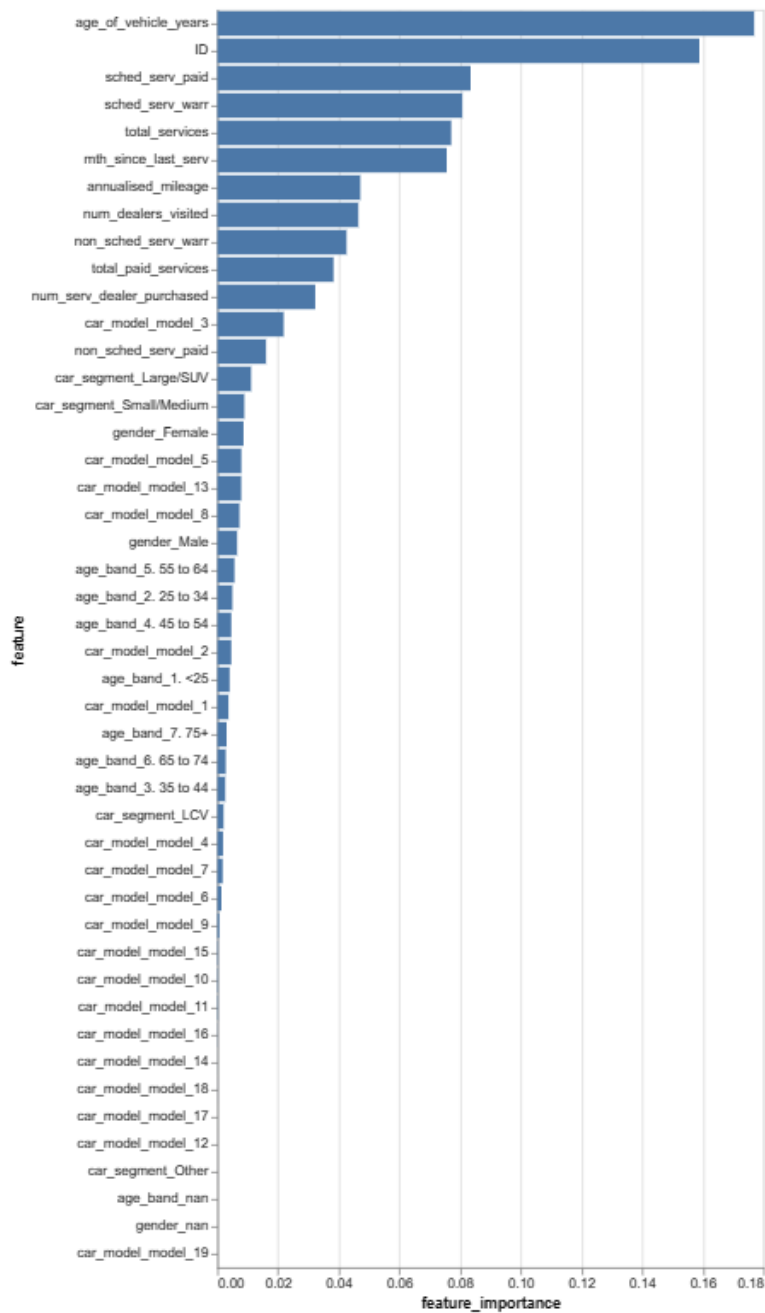




Figure 10: Feature Importance Chart

Data Preparation

Data preparation step for the classification machine learning models is an important step. Because the success and accuracy of predictions from the models deeply rely on this part. For training models with improper data would definitely lead the project to failure.

The following steps were taken in preparation of the raw data for it to be used:

- Data Loading and Exploration:
 - a. The data was loaded from a comma-separated values (CSV) file into a pandas DataFrame, a fundamental data structure in Python for data manipulation. The file name was “repurchase_dataset.csv”.
 - b. The dimensions of the dataset (number of rows and columns) were examined to understand its scope and the number of features available, including the identification of the target variable.
 - c. All features were explored to determine their data types, plots and initial statistics. This provided insights into data distribution and potential data quality issues.
- Data Cleaning
 - a. The dataset was checked for missing values. Rows containing missing data were removed to ensure models were trained on complete records only, preventing potential biases and errors.
 - b. Duplicate data was identified and eliminated to ensure a fair and accurate training process.
- Feature Engineering:
 - a. Categorical features (‘age_band’, ‘gender’, ‘car_model’, ‘car_segment’) were transformed using one-hot encoding. This technique creates binary columns for each category within a categorical feature, allowing machine learning algorithms to better understand and utilize the data for prediction. The original categorical feature was then removed from the dataset.

- 
- Data Splitting and Preparation:
 - a. The target variable was separated from the dataset to form the dependent variable (y). The remaining features became the independent variables (X). This separation is a crucial step for supervised learning tasks.
 - b. The dataset was then divided into training, validation, and testing sets. Initially, an 80/20 split was employed, allocating 80% for a combined training-validation set and 20% for the test set. Subsequently, the training-validation set was further split into 70% training data and 30% validation data. This multi-stage splitting ensures effective model training, validation for potential overfitting, and final testing to assess generalizability.
 - c. Feature scaling was applied to ensure all features have a similar range of values. This prevents models from being biased towards features with larger magnitudes and improves computational efficiency during the training process.
 - Handling Issues:
 - a. Rows with missing values were removed. While a straightforward approach, this might lead to data loss.
 - b. Also duplicate rows within the dataset were also removed.
- 

Modeling

Five machine learning algorithms were for this project. And they were chosen based on their strengths on our problem:

- **K-Nearest Neighbors (KNN):** A non-parametric classification algorithm that identifies the most similar data points from the training data to classify new data points.
- **Support Vector Machine (SVC):** A powerful classifier for high-dimensional data with clear separation between classes. It utilizes kernel functions to handle non-linear decision boundaries.
- **Decision Tree Classifier:** This algorithm creates a tree-like structure representing a series of decisions to classify data.
- **Random Forest Classifier:** An ensemble method that combines multiple decision trees for improved prediction accuracy and reduced overfitting.
- **Extra Trees Classifier:** Similar to Random Forest, but with increased randomization during tree generation, leading to high variance reduction.

Rationale for Selection:

- **KNN:** Selected for its effectiveness in classifying non-linear relationships. As KNN does not assume any underlying distribution for the data, it is suitable for real-world datasets that often do not adhere to theoretical distributions.
- **SVC:** Chosen for its ability to handle non-linear decision boundaries through various kernels, making it ideal for complex, high-dimensional classification tasks.
- **Decision Tree Classifier:** Preferred for its interpretability, allowing for clear understanding of the model's decision-making process.
- **Random Forest Classifier:** Utilized to leverage the power of ensemble learning to reduce overfitting while maintaining accuracy, particularly beneficial for large and high-dimensional datasets.
- **Extra Trees Classifier:** Incorporated for its strength in reducing variance, especially valuable for complex datasets with numerous features.

Hyperparameter Tuning:


Hyperparameter tuning was conducted to optimize the performance of each algorithm. Specific hyperparameters were adjusted for each model:

- Random Forest Classifier:
 - c. Hyperparameters Tuned: `n_estimators`, `max_depth`, `min_samples_leaf`.
 - d. Values Tested: `n_estimators` [200, 50], `max_depth` [15], `min_samples_leaf` [2, 10].
- KNN Classifier:
 - a. Hyperparameters Tuned: `n_neighbors`, `metric`.
 - b. Values Tested: `n_neighbors` [3, 10, 20, 40], `metric` [Euclidean, Manhattan].
- SVC
 - a. Hyperparameter Tuned: `C`.
 - b. Values Tested: `C` [2, 0.4, 0.05].
- Decision Tree Classifier:
 - a. Hyperparameter Tuned: `min_samples_split`.
 - b. Values Tested: `min_samples_split` [5, 10, 20, 30].
- Extra Trees Classifier:
 - a. Hyperparameters Tuned: `n_estimators`, `max_depth`, `min_samples_leaf`, `max_features`.
 - b. Method Used: Random Search CV to find the best values.

This process involved adjusting specific parameters to achieve a balance between model complexity and generalizability.

Random Forest and Extra Trees Classifiers: The tuning of these ensemble approaches was centered on variables such as the number of trees (`n_estimators`), the depth of the trees (`max_depth`), and the minimum number of samples per leaf (`min_samples_leaf`). The goal of these modifications was to achieve a balance between variance and model bias. Overfitting can occur from excessive variation, while underfitting can be caused by significant bias. By adjusting these settings, a strong ensemble of trees that could recognize the underlying patterns in the data without becoming unduly susceptible to particular training cases was produced.

Support Vector Machine (SVC): In SVC, the hyperparameter `C` regulates the trade-off between making sure the model generalizes well to new data and attaining a distinct separation between classes. By adjusting this parameter, it was possible to identify a sweet spot where the model could still perform well on fresh data while being able to distinguish between classes.



K-Nearest Neighbors (KNN): The distance metric and the number of neighbors, `n_neighbors`, taken into account, are the two most important hyperparameters for KNN. The model's overall accuracy and sensitivity to data points are determined by these parameters. By adjusting these hyperparameters, it was possible to identify the ideal distance metric to use when comparing data points and the ideal number of neighbors to take into account for classification.

Decision Tree Classifier: The minimal number of samples needed to split a node in the tree is determined by the `min_samples_split` parameter in decision trees. By ensuring that the tree doesn't grow unduly complicated and specialized to the training data, tuning this value helped to prevent overfitting. This is essential to preserving the interpretability of the model because simpler trees are typically easier to comprehend and explain.

■ ■ ■

Evaluation

a. Results and Analysis

After assessment has been carried out on the classification models for different tuned hyperparameters. We can see the following results:

1. Extra Trees Classifier:

- Default Hyperparameters:

Accuracy: Train 98.98%, Validation 98.95%, Test 98.83%

F1 Score: Train 0.0%, Validation 0.0%, Test 0.0%

Precision: Train 0.0%, Validation 0.0%, Test 0.0%

- Tuned Hyperparameters (Grid Search):

Accuracy: Train 98.98%, Validation 98.95%, Test 98.83%

F1 Score: Train 0.0%, Validation 0.0%, Test 0.0%

Precision: Train 0.0%, Validation 0.0%, Test 0.0%

2. Random Forest Classifier:

- For all hyperparameter tuning and also Grid Search Best Parameters:

Accuracy: Consistent 100% across all sets

F1 Score: Consistent 100% across all sets

Precision: Consistent 100% across all sets

3. Decision Tree Classifier:

- Default and for all hyperparameter tuning:

Accuracy: Consistent 100% across all sets

F1 Score: Consistent 100% across all sets

Precision: Consistent 100% across all sets

4. SVC:

- Default Hyperparameters:

Accuracy: Train 99.80%, Validation 99.52%, Test 99.67%

F1 Score: Train 89.80%, Validation 74.07%, Test 83.78%

Precision: Train 96.70%, Validation 85.71%, Test 100%

- For Hyperparameter C=2:

Accuracy: Train 99.88%, Validation 99.59%, Test 99.75%

F1 Score: Train 94.06%, Validation 79.55%, Test 88.61%

Precision: Train 97.94%, Validation 83.33%, Test 97.22%

- For Hyperparameters C=0.4:

Accuracy: Train: 99.40%, Validation set: 99.15%, Testing set: 99.26%

Precision: Train: 100% Validation: 84.61% Testing: 100%

F1 Scores: Train: 59.06% Validation: 37.28% Testing: 54.23%

- For Hyperparameters C=0.05:

Accuracy Scores: Training: 99.40% Validation 99.15% Testing 99.26%

Precision: Training 100% Validation 84.61% Testing 100%

F1 Scores: Training 59.06% Validation 37.28% Testing set: 54.23%

4. KNN Classifier:

- Default Hyperparameters:

Accuracy: Train 99.25%, Validation 99.07%, Test 99.18%

F1 Score: Train 44.60%, Validation 25.45%, Test 46.43%

Precision: Train 91.18%, Validation 77.78%, Test 100%

- 3 Neighbors & Euclidean:

Accuracy: Train 99.52%, Validation 99.16%, Test 99.24%

F1 Score: Train 69.57%, Validation 39.34%, Test 53.33%

Precision: Train 100%, Validation 80%, Test 94.12%

- 10 neighbors and Euclidean:

Accuracy: Training 99.11% Validation 99.06% Testing 98.93%

Precision: Training 100% Validation 100% Testing 100%

F1 Scores: Training 100% Validation 100% Testing 100 %

- 20 neighbors and Euclidean Distance:

Accuracy: Training: 99.01% Validation 98.97% Testing 98.90%

Precision: Training 100% Validation 100% Testing 100%

F1 Scores: Training 7.33% Validation 4.25% Testing 13.04%

- 20 neighbors and Manhattan Distance:

Accuracy: Training 99.00% Validation 98.97% Testing 98.93%

Precision: Training 100% Validation 100% Testing 100%

F1 Scores: Training set: 5.55% Validation 4.25% Testing 17.02%

- 40 neighbors and Manhattan Distance:

Accuracy: Training 98.97 Validation 98.95% Testing 98.82%

Precision: Training 100% Validation 100 % Testing 100%

F1 Scores: Training 0.0% Validation set: 0.0% Testing set: 0.0%

After going through all of the assessment upon the classification models with all types of hyperparameter tuning some key observations can be found:

1. Despite achieving high accuracy, the Extra Trees Classifier exhibited low F1 score and precision on the validation and test sets. This suggests potential bias towards the majority class, particularly if the dataset is imbalanced. The model might be simply classifying all instances as the majority class.
2. The Random Forest Classifier achieved perfect scores on all metrics across all data splits and hyperparameter configurations. While this may seem ideal, it is a strong indicator of overfitting. A well-generalizing model shouldn't achieve 100% accuracy on unseen data.
3. Similar to the Random Forest Classifier, the Decision Tree Classifier also exhibited perfect scores across all metrics. This again suggests overfitting to the training data.
4. The SVC demonstrated strong performance on the training set, with a slight decrease in performance on the validation and test sets. This is an expected behavior, as models typically perform better on data they are trained on. Notably, the SVC maintained a high F1 score and precision on the test set, particularly with a hyperparameter C value of 2. This suggests that SVC achieved better generalization compared to the tree-based models.
5. KNN's performance varied significantly depending on the chosen hyperparameter settings. With default settings, the F1 score and precision on the training set were considerably lower compared to the tree-based models. This might indicate that the model did not learn the training data patterns as effectively. However, the lack of perfect scores also suggests that KNN might not have overfitted as much as the other models, potentially leading to better generalization.

The exploration using the classification models: Extra Trees Classifier, Random Forest Classifier, Decision Tree Classifier, Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN), provided valuable insights into their performance and behavior.

1. The perfect scores achieved by both Random Forest and Decision Tree Classifiers strongly suggest overfitting. These models likely became too complex for the data, memorizing the training set rather than learning generalizable patterns.
2. The consistently high accuracy but low F1 score and precision of the Extra Trees model on the validation and test sets indicate a potential data imbalance. In such cases, accuracy is not a reliable performance measure.
3. Based on the analysis, SVC appears to be the best performing model in terms of generalizing to unseen data. The moderate and expected performance drop from the training set to the test set suggests a better fit for the overall dataset.
4. KNN's varying performance, particularly its lower F1 scores, suggests that it might be more sensitive to hyperparameter selection and the specific characteristics of the dataset. This highlights the importance of careful hyperparameter tuning for KNN models.


b. Business Impact and Benefits

Based on the analysis of the experiment results, the Support Vector Classifier (SVC) emerges as the most favorable candidate for this task. And with the help of SVC a very good impact will for the business.

- By identifying patterns in historical customer purchase data offers significant advantages for customer retention and engagement strategies:
 - By identifying customers with a high likelihood of repurchasing, businesses can focus efforts on retaining these valuable customers. This can be achieved through targeted outreach campaigns offering personalized incentives, leading to higher engagement and conversion rates.
 - By generating predictive insights that enable businesses to proactively address customer needs. This proactive approach can lead to increased customer satisfaction and loyalty, ultimately strengthening customer relationships.
- By analyzing past purchase data and customer preferences, the model can generate more accurate forecasts for specific vehicle features and models. This information allows businesses to optimize their inventory by stocking the right mix of vehicles, reducing the risk of both overstocking and stockouts. Improved inventory turnover can be achieved by ensuring the availability of in-demand vehicles while minimizing holding costs for slow-moving inventory.
- The model can help identify marketing campaigns with the highest potential return on investment (ROI). This allows businesses to allocate marketing resources more efficiently and maximize their marketing budget.

The SVC model helps to handle the challenges that other models struggle. And for this, using SVC to business needs would help grow the business exponentially.

- The model's robustness to high-dimensional data and ability to handle non-linear patterns might provide some level of resilience to data inconsistencies and missing values. With appropriate data preprocessing, the SVC model can still extract actionable insights even in the presence of complex data relationships.
- The SVC model exhibits adaptability thanks to the kernel trick, which allows it to learn new patterns without requiring explicit reprogramming. This is particularly beneficial for evolving datasets where customer behavior or market trends might change over time.
- Maintaining the SVC model's relevance and accuracy over time can be achieved through regular retraining and fine-tuning with updated data, ensuring its scalability and adaptability to market changes.



Implementing the SVC model can translate into quantifiable improvements for the business:

- **Increased Sales**

By predicting customer repurchases, businesses can tailor marketing campaigns more effectively. This can lead to significant increases in conversion rates, with even a 1% improvement potentially resulting in substantial revenue gains for large dealerships.

- **Reduced Marketing Costs**

Improved targeting through predictive analytics leads to more efficient marketing campaigns, potentially reducing customer acquisition costs and increasing marketing ROI. Businesses often report cost savings of up to 20-30% by utilizing predictive analytics for marketing optimization.

- **Inventory Cost Reduction**

Optimizing inventory based on demand forecasts generated by the SVC model can lead to significant cost reductions. Holding costs, which can account for 20-30% of the inventory's value, can be minimized by ensuring the availability of in-demand vehicles while avoiding overstocking of slow-moving inventory.

c. Data Privacy and Ethical Concerns

The use of customer data in predictive modeling raises significant data privacy and ethical concerns. And by proactively addressing data privacy and ethical concerns, businesses can ensure responsible and respectful data-driven initiatives that prioritize individual and community rights.

Privacy and Ethical Concerns:

- Customer data often includes demographics, purchase history, and potentially sensitive information. Mishandling of this data could lead to privacy violations.
- Customers should be informed about how their data is collected, used, and for what purposes. There must be a consent agreement.
- Only collect data essential for the analysis, avoiding unnecessary information that could infringe on privacy.
- Data or models reflecting societal biases can lead to unfair treatment of specific customer groups.

As there might be many concerns over the privacy and improper handling of data. Some measure might be taken by the business:

- Data anonymization by removing or encrypting personally identifiable information to ensure individual customers cannot be directly identified.
- Implement robust security measures to protect data from unauthorized access.
- Adhere to relevant data protection regulations regarding customer data handling.
- Be transparent with customers about data usage.

There must be no bias and always inclusiveness must be promoted by the business.

- Be aware of and sensitive to cultural differences in data interpretation, especially when dealing with Indigenous data that might have been historically misused.
- Ensure data practices don't exclude or misrepresent Indigenous people, leading to biased decision-making.
- Engage with Indigenous communities regarding data usage to understand their perspectives and concerns.
- Personal data might be misused for unintended purposes like segregation or discrimination. Implement strict access controls and regular audits to mitigate this risk.
- Being transparent in the model and data handling process to increase stakeholder understanding of decision-making processes, fostering trust and accountability.
- When using Indigenous data, consult with the respective communities to understand data usage implications and seek their input in model development and deployment.



Conclusion

This project yielded valuable insights into machine learning model performance and its alignment with business goals. The Support Vector Classifier (SVC) emerged as the most effective model. It achieved good performance on unseen data, suggesting its ability to generalize and make accurate predictions about customer repurchases. The high accuracy but low F1 score of the Extra Trees Classifier pointed to a significant data imbalance, a crucial factor to consider when evaluating model performance. The Random Forest and Decision Tree models suffered from overfitting, highlighting the importance of selecting models that can learn from the data without simply memorizing it.

The project successfully utilized machine learning and showed pathways to achieve key business goals if properly deployed:

- Enhancing customer retention through targeted engagement strategies based on predicted repurchasing behavior.
- Optimizing inventory management with improved demand forecasting based on customer preferences.
- Streamlining strategic marketing campaigns by allocating resources towards those indications with the highest predicted ROI.

The project would address stakeholder requirements by providing insights that can improve customer targeting, resource allocation, and overall customer satisfaction. This ultimately contributes to a better understanding of customer needs and improved business performance.

The future work would include two things deploying the model and making sure the data privacy and ethics are maintained.

- Implementing robust practices is crucial. This includes anonymizing data to protect customer privacy, acquiring informed consent for data collection and usage, and maintaining strong security measures to safeguard sensitive information.
- A comprehensive plan outlining the integration of the chosen model into the existing IT infrastructure should be formulated.
- A monitoring system should be established to track the model's performance in a production environment using real-world data. This will be accompanied by a maintenance schedule for periodically updating the model to maintain its effectiveness.
- Training materials and comprehensive documentation should be developed to educate end users on the model's functionality, its intended use within business processes, and its limitations.
- The deployment process should adhere to all relevant compliance regulations and data privacy rules, ensuring the integrity and security of customer data.

By prioritizing model optimization followed by a well-defined deployment plan, we aim to maximize the predictive power of the customer repurchase prediction model while guaranteeing a smooth and effective transition to real-world applications.



References

- Soofi, A., & Awan, A. (2017). Classification Techniques in Machine Learning: Applications and Issues. *Journal of Basic & Applied Sciences*, 13, 459–465. <https://doi.org/10.6000/1927-5129.2017.13.76>
- Cheriyan, S., Ibrahim, S., Mohanan, S., & Treesa, S. (2018, August 1). *Intelligent Sales Prediction Using Machine Learning Techniques*. IEEE Xplore. <https://doi.org/10.1109/icCECOME.2018.8659115>
- Gurnani, M., Korke, Y., Shah, P., Udmale, S., Sambhe, V., & Bhirud, S. (2017). Forecasting of sales by using fusion of machine learning techniques. *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*. <https://doi.org/10.1109/icdmai.2017.8073492>
- Amari, S., & Wu, S. (1999). Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6), 783–789. [https://doi.org/10.1016/s0893-6080\(99\)00032-5](https://doi.org/10.1016/s0893-6080(99)00032-5)

