Assignment 2Classification Models

short line

Rusan Vaidya

Student ID: 24886400  
26-05-2024

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)

[b. Key Objectives 2](#_kksv73ymqhus)

[**2. Data Understanding 4**](#_yg03y3w5hig9)

[**3. Data Preparation 8**](#_tj77rrmixzhi)

[**4. Modeling 10**](#_6vwh14nc44ga)

[**5. Evaluation 11**](#_mngiyx5tet6w)

[a. Results and Analysis 11](#_25v57zsj5m7g)

[b. Business Impact and Benefits 15](#_hkob76wu4d6q)

[c. Data Privacy and Ethical Concerns 16](#_uf1z6gbsejg6)

[**6. Conclusion 17**](#_9hr11g79asdx)

[**7. References 18**](#_ma3y8ytjvp1y)

# Business Understanding

## Business Use Cases

This project aims to develop a predictive model determining the likelihood of existing customers repurchasing a vehicle. Such a model is crucial for optimizing marketing strategies by identifying potential repeat customers who may respond well to targeted advertisements or promotional offers. Integrating this model into customer relationship management (CRM) systems can automate and personalize communications, enhancing customer engagement and retention strategies. Additionally, the insights gained from the model enable the sales team to focus their efforts on customers who are more likely to make repeat purchases, thus improving sales efficiency and effectiveness.

The main challenge in this research is the efficient allocation of marketing resources. A broad audience, including many individuals who may not be interested in purchasing a car, might be included as the marketing campaign’s target if they are not identified as potential repurchasing customers, leading to inefficient use of marketing budgets. However, the opportunity lies in capitalising on machine learning to decode the customer demographics, their first car attributes and purchasing behaviour to predict the purchasing likelihood of a second car based on historical data. By accurately predicting future buying behaviour, the company can create more effective and personalized marketing campaigns, likely leading to higher conversion rates and enhanced customer satisfaction.

## Key Objectives

This project is about developing a model to predict if customers will repurchase vehicles, aiming to focus marketing efforts, enhance customer service, and improve sales strategies effectively. The possible stakeholders include:

**Marketing Department**: They use the model to target campaigns better and manage budgets efficiently.

**Sales Teams**: They leverage insights to prioritize efforts and personalize pitches.

**Customer Service Representatives:** They improve service by understanding customer preferences and potential for future purchases.

**Management**: They evaluate project success and ROI to decide on future investments.

The diagram here visually represents how the predictive model connects with different stakeholders and the benefits involved, such as more targeted marketing, better customer retention, and higher sales conversions.





















short dash

# Data Understanding

The dataset provided for the predictive modelling project consists of various customer-related and vehicle service features formatted in deciles. This structured approach facilitates a more straightforward analysis and modelling process, as it standardises the scale across different variables.

One of the limitations of the dataset is the imbalance in the target variable, where a small fraction of the entries represent customers who have purchased more than one vehicle. This imbalance could lead to biased models towards the majority class, potentially overlooking subtler patterns distinct to the minority class. Additionally, the dataset has significant missing values in `age\_band` and `gender`, which could skew the model’s understanding of customer demographics unless properly addressed. Using deciles for all continuous variables while simplifying scaling might also mask underlying nuances in the data by reducing the granularity of information, potentially leading to a loss of detail that finer measurements could provide.

The dataset includes a variety of features that are pivotal for understanding customer behaviour and predicting repurchase probabilities:

1. **Demographic features:** age\_band` and `gender` help segment the customer base and tailor marketing strategies accordingly.
2. **Vehicle-attribute features:** `car\_model` and `car\_segment` provide insights into customer preferences for specific types of vehicles.
3. **Service utilisation features:** Variables like `sched\_serv\_warr`, `non\_sched\_serv\_warr`, `sched\_serv\_paid`, and `non\_sched\_serv\_paid` are crucial for analysing customer engagement with service offerings. High service use might indicate greater satisfaction or deeper engagement with the brand, potentially influencing repurchase decisions. Columns `mth\_since\_last\_serv`, `annualised\_mileage`, `num\_dealers\_visited`, and `num\_serv\_dealer\_purchased` can help identify patterns in service behaviour and loyalty to service providers, which are indicative of overall customer satisfaction and loyalty.

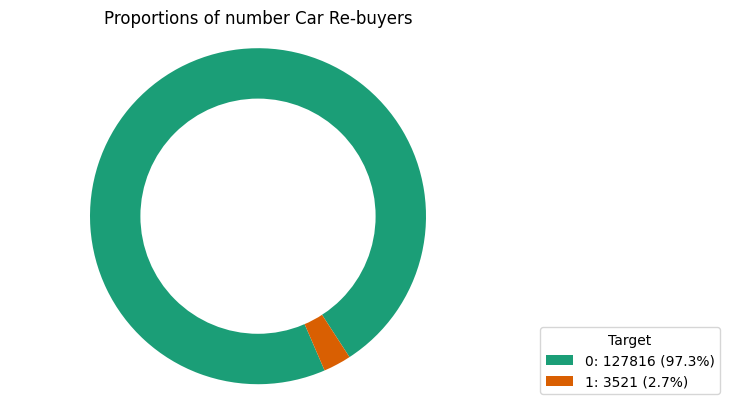


Figure: Proportion of number of customers who tend to buy a second car

From the plot above, we can figure that 97.3% of the individuals are not buying a new car, which suggests they might feel sufficient and happy with their first purchase.

2.7% of customers buy a new car, probably due to family growth, the expiration of an old car, or many other reasons.

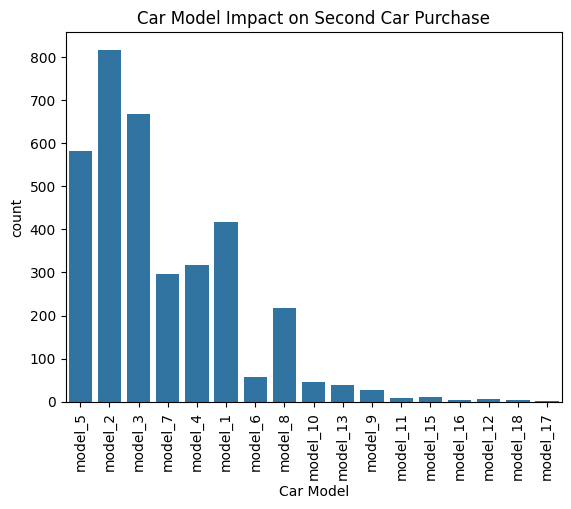
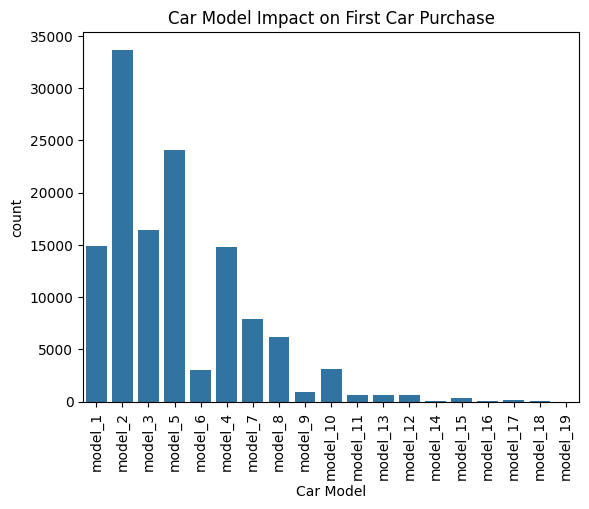


Figure: Car Model Impact on Car Purchase

We can observe that the Model 2 car is very famous, and most customers bought them as their first or second car. Many customers do not buy the newer models as they might rely on the older ones.

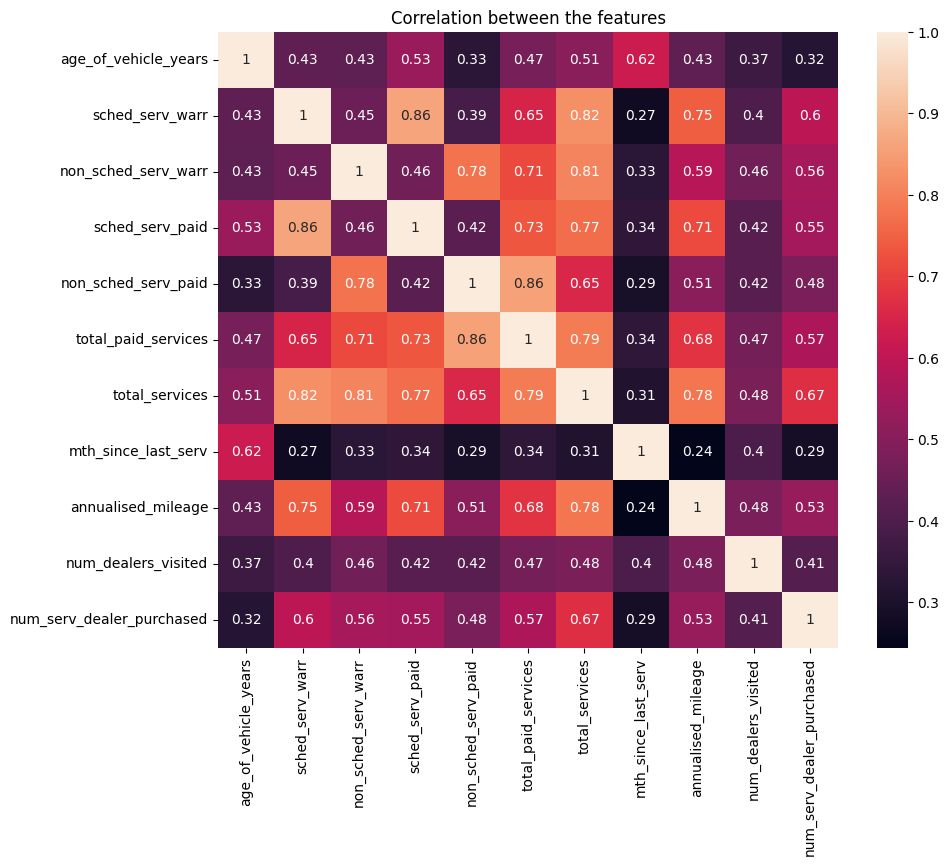


Figure: Correlation between car attributes

From the above figure, we can observe that scheduled services performed under warranty and non-scheduled services performed under warranty have a strong positive correlation with the age of the vehicle, which suggests that older vehicles have had warranty services performed on cars.

The other interesting thing found from this plot is that the last service has a negative correlation with the age of the vehicle, which might mean that older cars are coming to the end of their working span.

short dash

# Data Preparation

Data preparation ensures that the data is suitable for modelling. These steps typically include data cleaning, preprocessing, and feature engineering. Data preparation performed for this project is as follows:

**Dropping Identifiers**: Identifiers that do not contribute to predictive power, such as unique transaction IDs or user IDs, were removed.

**Experiment 1:**

1. **Handling Missing Values**: Columns containing null values were entirely dropped. This approach is straightforward but might lead to loss of valuable data if the dropped columns are informative.
2. **Mapping Categorical to Numeric**: Categorical variables were transformed into numeric form by mapping categories to integers. This method is useful for ordinal data but can introduce artificial ordinal relationships for nominal data.
3. **Handling Imbalanced Data**: **Random Under-Sampling**: This technique was used to address class imbalance by randomly reducing the size of the overrepresented class. While effective in balancing classes, it can lead to the loss of potentially useful data.

**Experiment 2: One-Hot Encoding and Oversampling**

1. **Data Preprocessing** **using** **One-Hot Encoding**: Unlike mapping, one-hot encoding was used to transform categorical variables into a binary matrix, preventing the introduction of artificial ordinality. This method increases the feature space and is suitable for nominal data.
2. **Handling Imbalanced Data** **using** **Over-Sampling**: To tackle class imbalance, over-sampling was employed to increase the size of underrepresented classes, potentially leading to better model performance but at the risk of overfitting due to repeated instances.

**Experiment 3: Feature Scaling with SMOTE**

1. **Scaling of Ordinal Categorical Values**: Ordinal categories were scaled appropriately to maintain the ordinal nature and distances between categories.
2. **Handling Imbalanced Data** **using** **SMOTE (Synthetic Minority Over-sampling Technique)**: SMOTE generated synthetic samples from the minority class. This approach helps by creating "synthetic" instances rather than over-sampling with replacement, thus providing more "general" models and mitigating overfitting to some extent.

**Experiments 4 & 5: Replication of Experiments 3 & 4 Settings**

* The same techniques were used in experiment 3, indicating a consistent approach to scaling. Since the ensembles and tree algorithms are unaffected by imbalanced data, Sampling techniques were not used.

short dash

# Modeling

In this predictive modeling project, a variety of machine learning algorithms were selected for their unique strengths in handling binary classification tasks. Logistic Regression provided a baseline due to its efficiency and interpretability, while Support Vector Classifier (SVC) was chosen for its adeptness in high-dimensional spaces. K-Neighbors Classifier, a non-parametric approach, leverages the majority label of nearest neighbors, useful for non-linear patterns. Decision Tree Classifier offers clear, graphical decision processes, and Random Forest Classifier enhances robustness against overfitting through an ensemble of decision trees.

Parameter tuning was essential, involving adjustments to regularization in Logistic Regression, class boundary margins in SVC, neighbor counts in K-Neighbors, and tree depths in Decision Tree and Random Forest. Random Forest was coupled with cross-validation to prevent overfitting and validate generalizability. Performance metrics like ROC-AUC and F1 score, precision, recall helped finalize model selection, aiming for a model that was both accurate and capable of generalizing to new data.

short dash

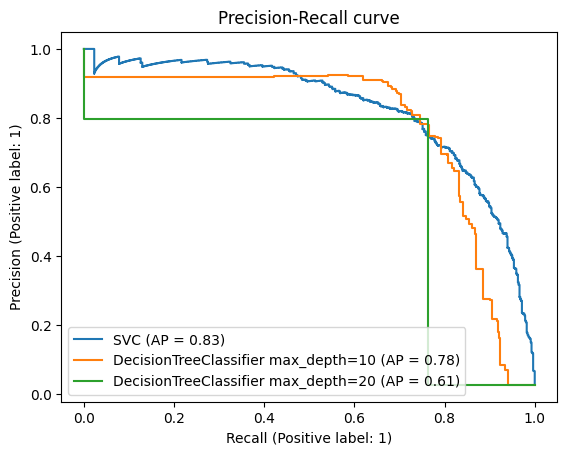
# Evaluation

## Results and Analysis

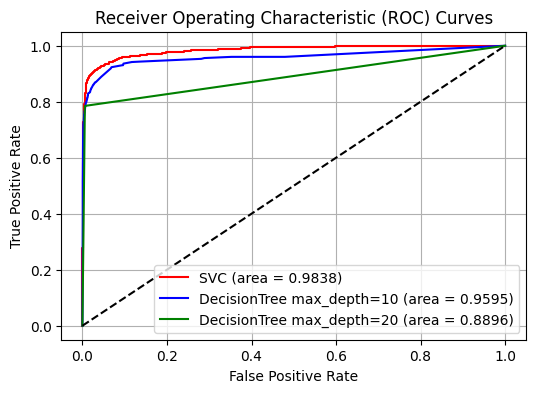
In the evaluation of machine learning models for predicting potential car buyers, the Support Vector Classifier (SVC) with C=1 showed a precision of 0.55 and a recall of 0.94, indicating it frequently correctly identifies potential buyers but also mistakenly classifies non-buyers as buyers. This could lead to unnecessary resource expenditure.

Meanwhile, Decision Tree Classifier performance varied with tree depth settings. At a max depth of 10, the precision was 0.95 and recall was 0.81, offering a more balanced accuracy, while a depth of 20 achieved perfect precision and recall, hinting at possible overfitting issues as such a model might not perform well with new, unseen data. The Decision Tree model, however, proved to be more reliable for practical business decisions due to its balanced accuracy and minimal error in identifying actual potential buyers. This balance helps conserve resources while effectively targeting likely customers. Insights from these experiments underline the importance of choosing and tuning models that not only achieve high accuracy but also align well with business objectives.

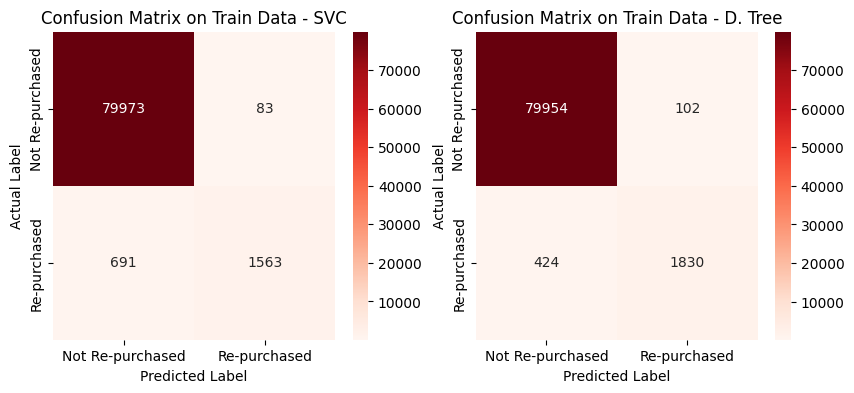
Ongoing refinement and testing of these models, especially the Decision Tree Classifier, are recommended to enhance prediction capabilities and support better decision-making in car repurchase scenarios.



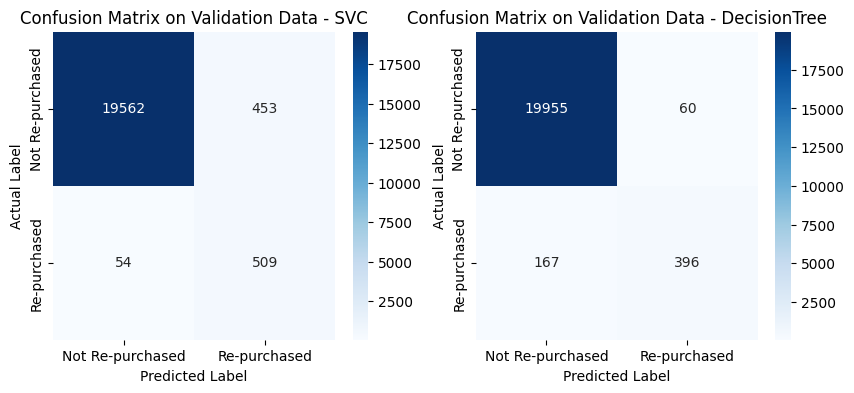
The Precision-Recall curve illustrates the trade-off between precision and recall for different threshold settings, with SVC outperforming both Decision Tree configurations, especially at higher recall levels.

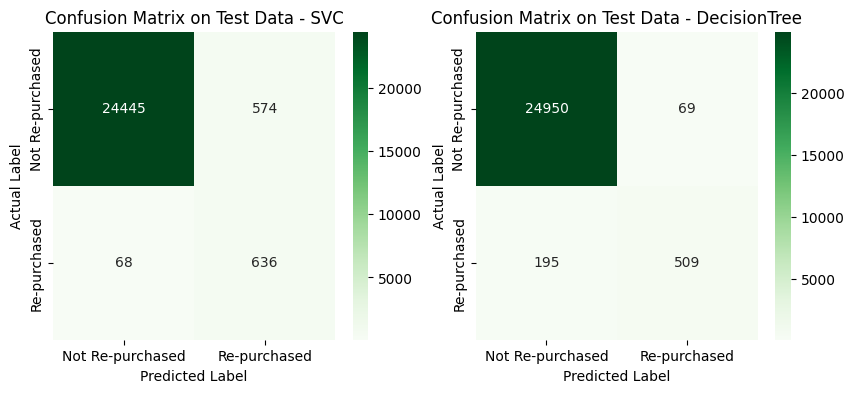


The ROC curve compares the true positive rate with the false positive rate at various thresholds, showing SVC as the most effective model, followed by the Decision Tree with max\_depth=10 and then max\_depth=20.



The confusion matrices for the SVC and Decision Tree models on training data show that the SVC model has more true negatives but fewer true positives than the Decision Tree, which suggests that the Decision Tree may be better at identifying re-purchases at the cost of slightly more false positives.





The confusion matrices indicate that on unseen data, the SVC model has a lower false positive rate compared to the Decision Tree, but the Decision Tree captures more true positives, which may make it a better model for identifying actual repurchases if the goal is to minimize missed opportunities.

## Business Impact and Benefits

The DecisionTreeClassifier model significantly enhances the capability to pinpoint potential car repurchase customers, leading to strategic advantages. By accurately identifying likely buyers, the model channels marketing efforts effectively, potentially boosting repurchases. This targeted approach not only increases revenue through higher sales conversions but also cuts costs by avoiding the pursuit of less interested customers.

Customer relations are also poised to benefit as precision in the model’s predictions translates to more personalized and relevant interactions, fostering loyalty. If each correct prediction adds to the profit margin, the amplified conversion rate directly impacts the bottom line positively.

The model's accurate targeting and customer engagement capabilities suggest a smarter allocation of resources, contributing to a more profitable and customer-centric operation. This progression towards data-driven decision-making stands to substantially elevate the business's financial performance and customer satisfaction.

## Data Privacy and Ethical Concerns

The project must adhere to stringent data privacy protocols to safeguard sensitive customer information. Ethical concerns arise around consent for data collection and potential biases in model deployment, which could inadvertently disadvantage Indigenous populations. To counteract these issues, robust anonymization techniques are employed, and models must be audited for fairness and bias. Regular reviews ensure compliance with legal standards must be conducted to understand any adverse effects on vulnerable groups, including Indigenous communities. These steps are essential to keep the project honest and ethical, making sure it treats every group of people fairly.

short dash

# Conclusion

The project supports DecisionTreeClassifier model to identify potential car repurchase customers with high precision and recall. The findings from this project include the model's balanced prediction capability, minimizing resource wastage while maximizing customer engagement opportunities. Insights reveal that targeted marketing based on the model’s predictions can enhance customer satisfaction and loyalty.

The model met stakeholders' requirements by improving the efficiency of marketing spend and potentially boosting repurchase rates. However, it's crucial to monitor the model for biases, especially regarding its impact on Indigenous communities and other vulnerable groups.

Future work should focus on refining the model with additional data over time to improve its predictive accuracy further. It's also recommended to explore additional features that could enhance the model's performance and to regularly review the model's fairness and privacy impact. Continuous updates and audits will help maintain the model's relevance and ethical application in dynamic market conditions.

short dash

# References

Imbalanced-learn.org. (2021). *imbalanced-learn documentation — Version 0.8.1*. [online] Available at: https://imbalanced-learn.org/stable/.

Scikit-Learn (2019). *User guide: contents — scikit-learn 0.22.1 documentation*. [online] Scikit-learn.org. Available at: <https://scikit-learn.org/stable/user_guide.html>.

scipy.org. (n.d.). *SciPy*. Available at: <https://scipy.org>.

www.rpubs.com. (n.d.). *RPubs - Document*. [online] Available at: https://www.rpubs.com/LexyBrooks/DamPartBDone [Accessed 26 Apr. 2024].

short dash