Assignment 2  
Classification Models

short line

Ngoc Quang Vinh Pham

Student ID: 25100660  
April 04th, 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

The project’s scenario assumes us as employees at a car seller company and focusing on building machine learning model to predict if an existing customer is more likely to re-buy a new car which will benefit the company to launch a marketing campaign targeting potential buyers.

Use case:

The sole use case for this project is to predict which customers are most likely to make further car purchases in the list of historical customer data that will help the business target them more successfully with specialized marketing campaign. This forward-thinking approach concentrates efforts on high-probability candidates in order to increase customer retention rates and maximize marketing expenses.

Additionally, enterprises can adjust inventory to align with the expected demand from repeat buyers, optimizing stock levels for specific models.

Moreover, by applying the project, we can increase the accuracy of financial planning and revenue forecasts by using predictive analytics to calculate the rate of client repurchases. This supports supply chain management, sales, and marketing professionals in making well-informed financial decisions.

Difficulty:

Nevertheless, the difficulties of effectively allocating marketing resources and optimizing sale opportunities in an extremely competitive car industry market are still concerning and it’s also the motivation of this project. Without predictive insights, traditional marketing techniques may lead to wasted resources on low-probability customers. Therefore, machine learning presents an opportunity since it can analyze large datasets and identify patterns that suggest a greater chance of consumer repurchase, which can be significantly useful when human analysis could be time-consuming, less accurate and could be done only by highly educated professional.

1. Key Objectives

The key objectives of the project:

* Enhance sales efficiency: The car businesses, by precisely anticipating the repurchase probability, can more effectively target potential repeat customers and raise the Return on Investment (ROI) for marketing campaigns.
* Resource Optimization: Allocating funds more wisely and concentrating on high-potential clients can assist you avoid spending money on unpromising prospects and optimizing budget allocations and resources.
* Boost customer engagement: More individualized customer experiences which encourage loyalty and raise customer satisfaction may result from personalized advertising that is based on predictive analytics.

Senior management, the sales department, and the marketing team are the project's stakeholders, and each has unique requirements:

* Marketing Team: To create focused campaigns and assess their success, they need predictive insights.
* Sales Department: To optimize sales procedures and concentrate efforts, it is necessary to identify prospective repeat clients.
* Senior Management: Seek overall improvements in ROI from marketing and sales initiatives as well as enhancements in sales efficiency.

In order to fulfill these objectives, the project will use machine learning algorithms to analyze consumer data and forecast purchasing patterns. This will help departments make more informed strategic choices and improve the operational effectiveness of marketing initiatives.

short dash

# Data Understanding

The project’s dataset comprises transactional records from a car dealership, tracking customer purchases and service history over several years. It consists of 131,337 records across 17 unique attributes, providing a thorough inspection of customer interactions and behaviors that might or might not influence vehicle repurchase likelihood.

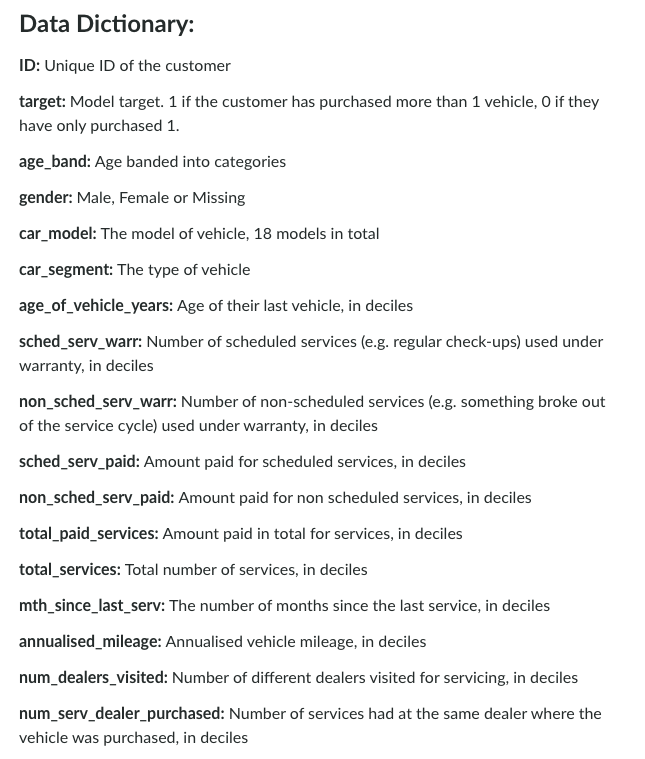
**Data Sources and Collection**

The dataset was provided from the project data sources and it is structured with various attributes reflecting customer demographics, vehicle details, and service interactions, essential for analyzing repurchase behaviors. Although the approach of gathering data aims for a thorough image, a few potential limitations are as follows:

**Data Limitation**

* **Missing and duplicated data:** There are only 2 attributes with a large proportion of null value which are ‘age\_band’ and ‘gender’. No duplicated values found in the dataset.
* **Bias in Vehicle Age**: Due to the decile-based age segmentation, automobiles that are much older or newer may not be as well-represented in the dataset, which might skew predictions in favor of the average age of the vehicle.
* **Service Location:** Due to geographic restrictions or a lack of alternatives, data on whether services are provided at the buying dealer may not accurately reflect customer preferences, which could bias the findings in favor of seeming loyalty.
* **Historical Data Application:** Longer-term trends or changes in consumer behavior, market circumstances, or economic variables that may have an impact on automobile buying and servicing patterns may not be taken into consideration if the dataset is limited to a certain period of time.

**Variables and Features Significance**

****

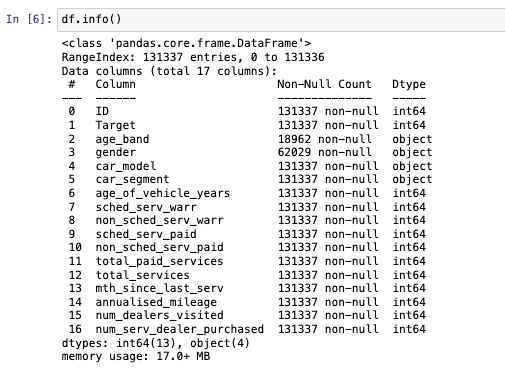
* De-identification data (ID, gender): Used to allow for targeted marketing campaigns and analysis while protecting client privacy.
* Car model and Segment: These are important variables in inventory management and targeted marketing as they affect consumer preferences and service requirements.
* Vehicle Age: Based on the vehicle lifespan, this indicates the possibility of servicing requirements and the possibility of repurchasing.
* Service History (Scheduled and Non-Scheduled): The service department's planning and client retention methods heavily rely on the service history which reflects vehicle dependability and customer loyalty.
* Financial Transactions (Amount Paid for Services): Offers information on consumer spending patterns and service businesses' profitability.
* Service Frequency and Dealer Visits: Indicators of customer engagement and satisfaction, servicing frequency and dealer visits are essential for improving loyalty programs and customer service.
* Annual Mileage: This may assist customize sales offers and servicing reminders by revealing trends of vehicle use.
* Last Service Interval: Assists in forecasting upcoming service appointments and client interaction possibilities.

**Exploratory Data Analysis (EDA)**

The dataset contains 131337 rows of record with 16 columns of features and 1 column is the Target variable.



As expected, there are 2 columns consist of an abundance amount of null value which is ‘age\_band’ and ‘gender’ which can be dealt with in the data preparation process. However, there is no duplicate value identify in the dataset.



* Numerical Features: The dataset contains 12 numerical features that cover various aspects of customer interactions and transactions, such as the age of vehicles, costs associated with services, and frequencies of service usage.
* Categorical Features: There are four categorical features:
  + age\_band: Groups customers into different age categories.
  + gender: Categorizes data by gender.
  + car\_model: Indicates the model of the car purchased or serviced.
  + car\_segment: Classifies the type of vehicle.

A blue line on a white background

Description automatically generated

The distribution of the target variable is strongly biased in towards Class 0. Machine learning models may become biased in favor of the majority class as a result of this imbalance. Before training models, it is crucial to apply resampling strategies to ensure that the models do not overlook the minority class.

* age\_band: 85% missing values.
* gender: 60% missing values.

Impact of Missing Data:

* Age Band: A large proportion of missing values in the 'age\_band' column could seriously harm the model's ability to estimate how age affects the chance that a vehicle will be repurchased. A degree of insufficient information like this might make any imputation technique inaccurate.
* Gender: Missing values in the 'gender' column may result in the loss of demographic data that might be useful in understanding purchase patterns. Hence, I decided not to present the distribution of the distribution.

A close-up of a graph

Description automatically generated

The distribution of 'car\_model' seems to be right-skewed, with the maximum frequency seen in the first category, which is probably'model\_1'. 'Model\_1' seems to be the most popular automobile model in this dataset, as seen by the progressive drop from'model\_2' to'model\_19'.

It shows three categories on the 'car\_segment' side. 'Luxury' and'small/medium' are the first two groups, and they may be the most common in the statistics because of their comparatively comparable high frequency. 'LCV' or 'Other' might be the last group. Its frequency is much lower, indicating that this kind of automobile segment is not as prevalent in the dataset.

**Correlation Matrix**

**A green squares with black lines

Description automatically generated with medium confidence**

**A screenshot of a computer code

Description automatically generated**

Correlation values close to 1 or -1 indicate a strong positive or negative linear relationship, respectively. Values close to 0 suggest no linear correlation. These insights can guide feature selection for predictive modeling, as features with higher absolute correlation may have more predictive power regarding the target.

The dataset's correlation analysis shows that, in contrast to widely accepted criteria like car model frequency and car segment type, a customer's gender may have a minimal to weak effect on their chance of buying several vehicles. There is a weak to moderate negative association between warranty utilization and service-related expenditures, indicating that consumers who suffer greater service costs are less likely to make future purchases. Age of the vehicle, number of services performed, and time since last service all have a negative correlation with the likelihood that a client would repurchase, suggesting that customers with older vehicles or more maintenance requirements may be less inclined to do so. There seems to be a modest negative correlation between the aim and visits to various dealers, services rendered at the dealer of purchase, and greater yearly miles. This correlation may be attributed to customer loyalty and use habits.

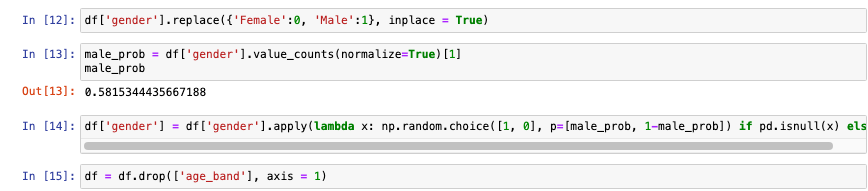
short dash

# Data Preparation

**Data Cleaning**

Handling Null value: Features that have more than 60% null values will be removed from the dataset due to the possibility of inaccurate and incorrect imputation of these characteristics.

* The gender variable was encoded as binary (1 for Male, 0 for Female) to facilitate modeling. Missing gender data was imputed to reflect the existing distribution, with 58% Male.
* The 'age\_band' feature was dropped due to a high volume of missing values (85%), which could lead to significant prediction errors if imputed incorrectly.

****

**Data Preprocessing and Feature Engineering**

* In order to convert categorical data into binary variables and prevent ordinal implications while maintaining discrete categories, the four unique values of "car\_segment" were one-hot encoded.
* Frequency encoding was used for 'car\_model' in order to minimize dimensionality and address its high correlation with the goal, effectively reflecting the popularity of car models.
  + Pros: Converts categories into a single numerical feature, preserving the importance of each category in terms of its frequency.
  + Cons: Different models with the same frequency are indistinguishable, potentially losing some information.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* Feature Selection: Since the large amount of the dataset, it’s best to consider all possible data in the machine learning algorithms for the refinement of the model. Therefore, I have consisted of all the attributes into the model expect the Target variable and the ‘ID’ identification data which will help signifying the likelihood of a customer making another vehicle purchase.

A close-up of a computer code

Description automatically generated

* Feature Scaling: I have presented an experiment to examine which scaling methodology is best for the model performance and Standard Scaler ranks one of the highest and was chosen for normalization process of this project. Scaling Method was chosen for the assessment are: MinMax Scaler, MaxAbs Scaler, Standard Scaler, and Robust Scaler.

**A screenshot of a computer

Description automatically generated**

**Data Splitting**

By applying train\_test\_split from library sklearn, I have split the data into 3 different datasets:

* Training dataset: 91935 records (70%)
* Validation dataset: 19701 records (15%)
* Testing dataset: 19701 records (15%)

A screenshot of a computer

Description automatically generated

The Standard SMOTE technique is implemented in our training dataset in order to enhance the complexity of the models by synthetically over-sampling minority data and then apply the trained model to the unseen validation and testing dataset with a highly imbalanced distribution. The selection of the SMOTE technique was also be examined by one of the project experiments where we were assessing 6 different techniques on different machine learning models in order to find what’s best for the predictive analysis outcome and further be applied in the real industry.

Various SMOTE techniques tested include:

* Standard SMOTE
* Borderline SMOTE
* SVM SMOTE
* ADASYN
* SMOTEENN
* SMOTETomek

A screenshot of a computer

Description automatically generated

short dash

# Modeling

**Experiment 1: Machine Learning Model Comparison and Evaluation Metrics Selection**

Conventional accuracy measures have been replaced to the benefit of more insightful measures like the F1 Score and ROC-AUC in order to address the imbalance in the data set. An F1 score of zero was obtained for all datasets when the baseline model was created using the most frequent class as a prediction. This was a clear indication that more advanced modeling approaches were required.

* F1 Score: The weighted average of Precision and Recall. This is useful when you seek a balance between Precision and Recall.
* ROC-AUC: Area Under the Receiver Operating Characteristic Curve. A higher AUC indicates a better model performance.

A screenshot of a computer program

Description automatically generated

For the first experiment, we tend to evaluate 6 machine learning models using Kfold from sklearn.model\_selection library to compare their performance in predicting a binary outcome using F1 Score and ROC-AUC as the evaluation metrics. The machine learning models are: Logistic Regression, Decision Tree Classifier, Random Forest Classifier, XGBoost Classifier, Knearest Neighbour Classifier, Support Vector Machine.

At the same time, we are examining which scaling technique is best for the performance of the model which included 4 different techniques: MinMax Scaler, MaxAbs Scaler, Standard Scaler, Robust Scaler.

**Experiment 2: Synthetic Over-Sampling Technique Application**

The top three models (Random Forest, XGBoost, and Decision Tree) from Experiment 1 were subjected to SMOTE methods. Standard SMOTE, Borderline SMOTE, SVM SMOTE, ADASYN, SMOTEENN, and SMOTETomek were among the many SMOTE techniques that were evaluated. Examining how various over-sampling strategies affected the models' capacity for outcome prediction was the aim. Each model was evaluated using the same default hyperparameters as in the previous experiment, focusing on tuning aspects directly affected by class distribution changes due to SMOTE.

SMOTEENN (SMOTE with Edited Nearest Neighbors)



SMOTE (Original)



BorderlineSMOTE



SVMSMOTE



ADASYN (Adaptive Synthetic Sampling)



SMOTETomek



**Experiment 3: Evaluation Metric Assessment with F-Beta and ROC-AUC**

Extreme Gradient Boosting Classifier, Random Forest Classifier, Support Vector Machine, and K-Nearest Neighbors were the four top-performing models from the previous experiments that were selected. To find out how well each model handled precision-recall trade-offs, these were then analyzed using F1, F2, F3, and ROC-AUC scores. All metrics demonstrated the excellence of the XGBoost model, including its high sensitivity to recall-focused measures and its ability to strike a balance between accuracy and recall.

The F1, F2, F3, and ROC-AUC scores were used to assess each model's ability to manage the trade-off between precision and recall:

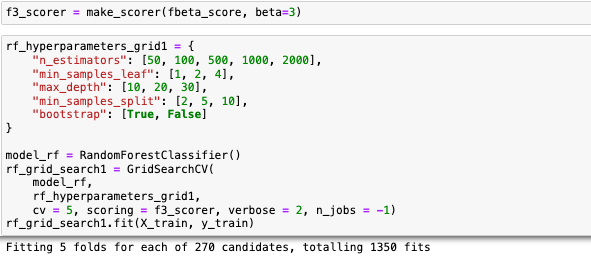
* F1-score: evaluate the balance between precision and recall.
* F2-score: put more weight on recall, useful in scenarios where missing a positive instance has greater consequences.
* F3-score: give much more weight to the recall making it appropriate for highly sensitive cases.
* ROC-AUC-score: measure the model’s capacity to distinguish between classes at different threshold values.

**Experiment 4: Random Forest Hyperparameter Tuning**

Random Forest was chosen because it can handle a wide range of data types and is resistant to overfitting, which is especially important when dealing with unbalanced datasets like our own. The grid search investigated a wide range of hyperparameter, including:

* The number of trees (n\_estimators) was evaluated between 50 and 2000. While adding additional trees usually increases model accuracy, doing so increases processing overhead.
* Tree Depth (max depth): To find the ideal level that captures sufficient data information without going into overfitting, we investigated depths of 10, 20, and 30.
* The numbers we select for splits at 2, 5, 10 and for leaf nodes at 1, 2, 4 are the Minimum Samples per Split (min\_samples\_split) and Leaf (min\_samples\_leaf). These factors aid in controlling the trees' development, making sure they are functional but not too precise to the point where they mimic the noise in the data.
* Bootstrap: We conducted experiments using (True) and (False) bootstrap samples. This decision has an impact on sample selection for tree construction, which in turn influences the bias and variance of the model.

The F3-score was chosen as the main assessment statistic for this tuning procedure. Recall is prioritized above accuracy by the F3-score, which is important in our situation since spotting possible recurring clients is more important than preventing false positives. This measure ensures that all possible positive occurrences (repurchases) are accurately identified by our algorithm, which aligns with our business goal of optimizing the efficacy and impact of marketing.



**Experiment 5: Extreme Gradient Boosting Hyperparameter Tuning**

For this experiment, the Extreme Gradient Boosting (XGBoost) model was used because to its robustness in classification tasks, especially when dealing with complex and imbalanced datasets. XGBoost is well recognized for its effectiveness and efficiency in managing many types of data and its ability to prevent overfitting, thanks to its extensive variety of hyperparameters for customization. We completely optimized some key parameters using grid search in order to fine-tune the model's performance:

The steps in the tuning procedure were:

Step 1: To create a baseline, train using the default parameters.

Step 2: To prevent overfitting, optimize "max\_depth" and "min\_child\_weight."

Step 3: Managing regularization by adjusting "gamma".

Step 4: Modifying "colsample\_bytree" and "subsample" to improve model performance.

Step 5: Adding an additional layer of complexity control in step five involves adjusting the regularization parameter 'lambda'.

Step 6: To more successfully converge to the optimum solution, make final tweaks to the 'n\_estimators' and lower the 'learning\_rate'.

The grid search investigated a wide range of hyperparameter, including:

* n estimators: Increased to 2000 at a higher computational expense to improve the flexibility of the model.
* max\_depth: Set at 5 to balance model complexity and prevent overfitting.
* min\_child\_weight: Set to 1 to control how sensitively the model responds to data changes.
* Gamma: Set at 0.5 to control the formation of leaf nodes and provide regularization.
* subsample and colsample\_bytree: By randomly selecting data points and features, both parameters are set at 0.9 to minimize overfitting.
* reg\_alpha: To select features, a little L1 regularization term of 0.001 was used.
* objective: For binary classification jobs, configured as "binary:logistic".
* scale\_pos\_weight: Maintained at 1 to maintain class weight parity.
* seed: For repeatability, set at 27.

The F3-score was chosen as the main assessment statistic for this tuning procedure. Recall is prioritized above accuracy by the F3-score, which is important in our situation since spotting possible recurring clients is more important than preventing false positives. This measure ensures that all possible positive occurrences (repurchases) are accurately identified by our algorithm, which aligns with our business goal of optimizing the efficacy and impact of marketing.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

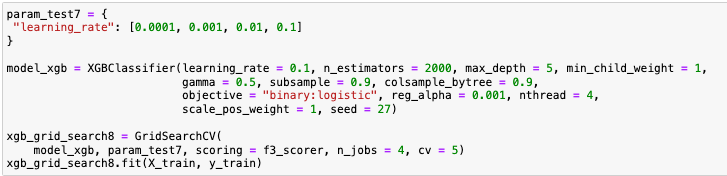
Description automatically generated

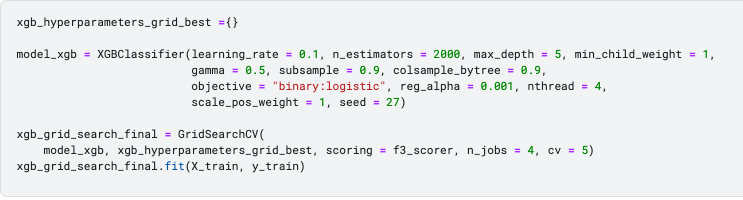
A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated





short dash

# Evaluation

## Results and Analysis

The model assessments used a systematic scaling method, generated synthetic data using SMOTE methods, and prioritized the F3-score above other performance indicators.

**Experiment 1**

It was important to pick a scaling approach, and the Standard Scaler was chosen since it was good at normalizing the feature set, which helped with better model convergence and generalization.

A screenshot of a computer program

Description automatically generated

For this project, however, I think the Standard Scaler—which received an F1-score of 0.845—is the better option. Based on its capacity to remove the mean and scale to unit variance, which typically guarantees that the existence of outliers does not bias the model's performance, this choice was selected.

We’re also using cross-evaluation on 6 different machine learning model and calculating the f1-score and standard deviation:

LR\_model:0.3184040754354613(0.015763525652172008)

DT\_model:0.7769255304038888(0.015342083339219652)

RF\_model:0.8524664719593253(0.006907857712558807)

XGB\_model:0.8740457966092746(0.011865063531119515)

KNN\_model:0.6241727581242291(0.01290486624554512)

SVM\_model:0.6736390759206159(0.016083993838863386)

A diagram of a graph

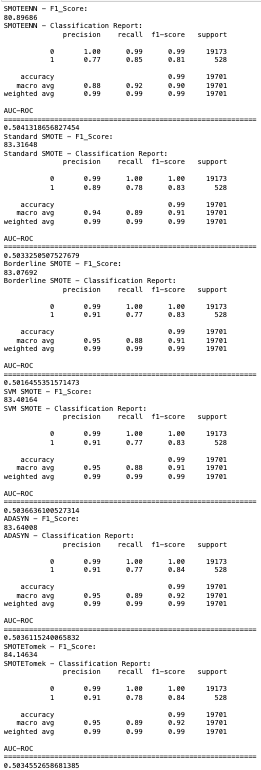
Description automatically generated

**Experiment 2**

Different strategies were evaluated throughout the SMOTE application in order to rectify the class imbalance present in the dataset. The impact of each SMOTE technique on the model's minority class classification performance varied:

* The best F1-scores were obtained by Standard SMOTE and SMOTETomek, suggesting a superior trade-off between accuracy and recall.
* Additionally, the F1-scores of ADASYN and SVM SMOTE increased, improving the sensitivity and specificity of the model.
* Although they showed modest gains, borderline SMOTE and SMOTEENN outperformed the others by a little margin.

**Random Forest Model**



**Extreme Gradient Boosting Model**

A screenshot of a document

Description automatically generated

**Decision Tree Model**

A screenshot of a document

Description automatically generated

In the training set, a fairer representation of the minority class was made possible by the use of Standard SMOTE, which was shown to improve the balance between classes. This resulted in better model sensitivity towards prospective repurchase cases.

AUC-ROC values were close to 0.50, indicating that while the F1-score increased, the models' capacity to distinguish between classes at various thresholds did not show a statistically significant improvement.

**Experiment 3**

The XGB model demonstrated superior technical performance across all metrics, indicating a robust capability to balance precision and recall and adjust to a recall-oriented focus (F2, F3). Random Forest followed closely, excelling particularly in ROC-AUC, suggesting strong discrimination abilities. KNN underperformed across the board, suggesting it may be less suited for the data's complexities. SVM showed adequate discrimination but lower precision-recall balance.

**F1-Score**

A graph of different colored bars

Description automatically generated

**F2-Score**

A graph of different colored bars

Description automatically generated

**F3-Score**

A graph of different colored bars

Description automatically generated

**ROC-AUC-Score**

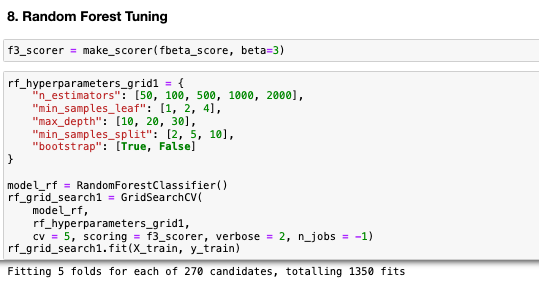
A chart of different colored bars

Description automatically generated

Because the F3-score puts more importance on recall than precision, it was crucial to the evaluation. Since it was determined that the cost of missing a genuine positive—a repeat purchase—was higher than the cost of false positives, this metric was especially appropriate for the company's objective of identifying repeat consumers.

**Experiment 4**

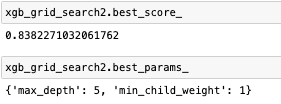
The tuning of the Random Forest model markedly improved its capability to identify likely repeat customers, as evidenced by the increase in the F3-score from 76.2 to 79.55. This score enhancement indicates a better recognition of minority classes, which is crucial in scenarios where the cost of missing a potential repeat customer is high. The optimal parameters—no bootstrapping, a maximum depth of 30, only one sample per leaf, a minimum of two samples required to split a node, and 500 trees—helped the model learn more nuanced patterns in the data without overfitting. This depth allowed the Random Forest to capture intricate structures in the data, enhancing its ability to predict customer behavior accurately.



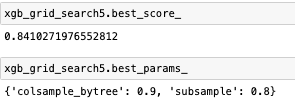
**Experiment 5**

The adjustment of the XGBoost model's hyperparameters led to a noteworthy improvement in its performance, with the F3-score rising from 82.247 to 84.391. This increase underscores the effectiveness of the tuned parameters in enhancing model accuracy, particularly in terms of recall. The choice of the F3-score as a metric was strategic, emphasizing recall to ensure the model effectively identifies potential repurchases. This focus aligns closely with the business objective of increasing car repurchase rates, demonstrating the model's enhanced capability to detect high-value customers likely to repurchase.

A screenshot of a computer code

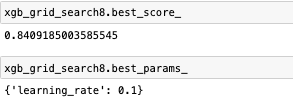
Description automatically generated

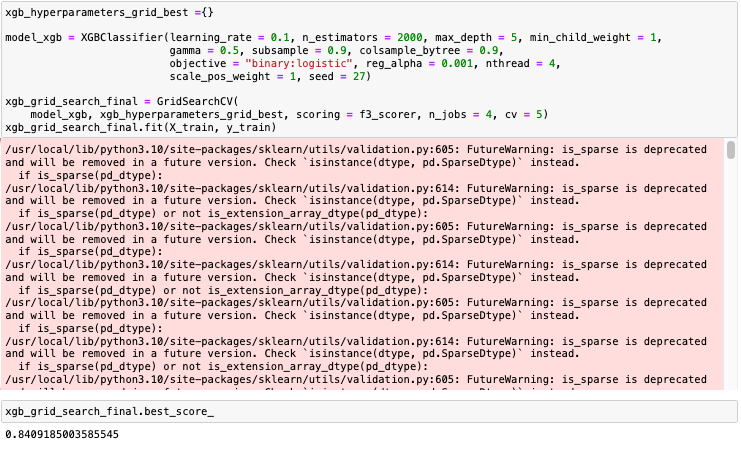
A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generatedA screenshot of a computer code

Description automatically generated



## Business Impact and Benefits

## Impact on Business Use Cases:

## The predictive model enhances targeted marketing strategies within the automotive sales industry by identifying customers most likely to repurchase vehicles. This capability allows the business to focus marketing efforts more efficiently, potentially increasing conversion rates and customer retention. The final model's improved predictive accuracy has a substantial impact on the company by enabling more effective resource allocation and enhancing marketing efforts. It allows for more precise and effective targeting of marketing interventions by identifying customers who are most likely to make additional purchases. This precision minimizes wasted efforts on less likely prospects and maximizes the return on investment.

## Exploiting Opportunities:

## The model opens up opportunities for personalized marketing by predicting customer repurchase probabilities. Tailored promotions and offers can be crafted to meet the specific needs and preferences of identified potential repeat customers, thereby enhancing customer engagement and satisfaction. For example, the improved F3-score in the XGBoost model (from 82.247 to 84.391) signifies a higher capability to identify potential repeat buyers. This enhancement ensures that marketing initiatives are not only more cost-effective but also yield higher returns by focusing on those with a history of multiple purchases.

Implementing the methodology is expected to raise repurchase rates, potentially leading to considerable revenue growth. If the plan is successful, even a 1% increase in targeted customer repurchase rates might provide significant financial gains given the average car selling price. The model also reduces false negatives, or missed chances, which boosts income. This method optimizes automobile company operations by improving marketing campaign efficiency and strategic financial planning and forecasting.

## Data Privacy and Ethical Concerns

Managing sensitive and personal data requires careful attention to data privacy and ethical issues. The project tackles these elements as follows:

Data Privacy Implication:

* Data Collection: It may be unethical to gather data without the persons' express agreement. However, the data was provided for us from car seller company that already get permission from the customers about the types of data being collected and how they will be used.
* Data Usage: It is unethical to use data for reasons other than those that the customers, who are the data subjects, have consented upon. Only business processes directly connected to the services the dealership offers should benefit from the utilization of the data.
* Model Deployment: If predictive models are not closely watched, they may be used for discriminatory or profiling purposes. The model results must be regularly verified to make sure no specific group of people—including indigenous communities that may be overlooked in the data—is mistakenly affected.

Data Privacy and Ethical Handling:

* Anonymization of Data: To eliminate or hide personally identifiable information (PII) data should, wherever feasible, be anonymized. This reduces the possibility of privacy violations.
* Secure Data Storage and Access Controls: Implementing strong security measures and limiting data access to authorized workers are two ways to ensure secure data storage and access controls.
* Impact assessments: Conducting routine analyses to identify and eliminate any biases or adverse effects on communities who are already at risk.

By addressing these concerns and implementing robust privacy protection measures, the project can ensure that its benefits are realized without compromising the ethical standards and privacy of the individuals involved.

short dash

# Conclusion

**Key Findings and Insights:**

The research demonstrated the effectiveness of sophisticated machine learning approaches in improving predictive analytics in the car sales business. Ensemble techniques, SMOTE, and intensive hyperparameter tuning have increased model performance and predicted accuracy. The experiments demonstrated that XGBoost, in particular, outperformed other models, showcasing its robustness in handling imbalanced datasets and complex predictive tasks. The main conclusions include the need of customized feature engineering and carefully considered model selection to meet certain business requirements like targeted marketing and client retention.

**Project Achievement:**

By creating a model, particularly the XGBoost model, that precisely anticipates the chance of a consumer making another purchase, the project succeeded in accomplishing its goals. This has made it possible to implement resource allocation methods, marketing plans, and sales processes that are more effective. Stakeholders in marketing, sales, and senior management now have better tools for identifying possible repeat customers thanks to the XGBoost model's efficacy in forecasting consumer repurchase probability. This has enhanced marketing efforts and increased return on investment.

**Future Work and Recommendations:**

* Model Improvement: To keep models accurate and relevant, models must be updated and trained on a regular basis due to the changing nature of consumer behavior and market situations.
* Model Exploration: Building on the success of the XGBoost model, further research into other machine learning algorithms may provide more insightful results and even better performance.
* Operational Deployment and Monitoring: To guarantee continuous performance and enable fast upgrades, implementing these models into real-world business environments should be supported by thorough monitoring.

The project's success opens the door for the adoption of its approaches in other areas of the company, offering a model for using advanced analytics to support operational effectiveness and business strategy. These initiatives have the potential to greatly improve a company's ability to adjust to changing marketplace dynamics and client demands, which might eventually result in an unchanging edge in competition.

short dash

# References

<https://pranav-c.medium.com/smote-vs-smote-enn-which-is-more-effective-for-churn-prediction-in-imbalanced-banking-data-b289414366a0>

<https://mriquestions.com/frequency-encoding.html#:~:text=Frequency%2Dencoding%20of%20spatial%20position,as%20a%20function%20of%20position>

<https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222>

<https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/>

<https://realpython.com/python-counter/>

<https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/>

<https://scikit-learn.org/stable/modules/cross_validation.html>

<https://machinelearningcoban.com/2017/08/31/evaluation/>

short dash