Program Profile		
Program	Program name	Optimizing Car Sales: Predictive Modeling for Automotive Customer Retention
	Category	A3

Summary of Program	
Program Name	Optimizing Car Sales: Predictive Modeling for Automotive Customer Retention
Category	A3
Abstract of Program	This project aims to improve customer retention and sales efficiency for an automotive company using advanced machine learning techniques. The study analyzes a comprehensive dataset of over 130,000 customer records, encompassing car purchase history, service interactions, and engagement metrics, to predict which customers are most likely to repurchase vehicles. A major challenge addressed is the significant class imbalance, as only 2.68% of customers have made a repurchase, necessitating robust modeling and evaluation strategies. The data preparation phase involved handling missing values, selecting relevant features, and encoding categorical variables. Multiple classification algorithms were tested, including Support Vector Machines, Decision Trees, Random Forests, and Extra Trees. To manage the imbalanced data, model evaluation emphasized the ROC AUC score, balancing sensitivity and specificity. Among the models, the Decision Tree with default parameters achieved the best initial results, producing a test ROC AUC of 0.84. This model was chosen for its balance of predictive performance, interpretability, and computational efficiency. It is important to note that the current results are based on a demo dataset provided by stakeholders, and the system remains in development. Future work includes integrating the predictive model into operational workflows, enhancing feature engineering, performing hyperparameter tuning, and further addressing data imbalance to increase robustness. These steps aim to create a reliable tool for identifying high-value customers, thereby supporting data-driven decision-making in marketing and sales strategies. Overall, this project demonstrates the potential of machine learning to optimize customer retention strategies in the automotive sector by providing actionable insights into repurchase behavior, while highlighting the importance of careful handling of imbalanced datasets for predictive modeling.

Details of Program		
Planning		
Objectives	Long-term Goals	• Develop a scalable predictive analytics platform for automotive customer retention that can be deployed across multiple dealerships and brands. • Enhance data-driven decision-making in sales, marketing, and

		inventory management by continuously improving predictive models with real-world customer data. • Incorporate advanced machine learning techniques such as ensemble methods, deep learning, and Explainable AI (XAI) to improve model accuracy and transparency. • Foster industry-academic collaboration by integrating research outputs into practical business solutions, providing insights that influence marketing strategy and customer engagement practices. • Promote student and faculty expertise in applied machine learning, business analytics, and AI-driven customer relationship management, positioning WUB as a leader in innovative educational programs.
	Short-term Targets	• Transition from demo to live datasets from automotive clients with proper anonymization and compliance protocols. • Refine existing machine learning models (Decision Tree, Random Forest, SVM, Extra Trees) through hyperparameter tuning and feature engineering. • Evaluate model performance using ROC AUC, precision, recall, and F1-score, ensuring accurate detection of high-value customers. • Pilot deployment of the predictive model within a small customer segment to test operational integration and real-time impact. • Train students and faculty in advanced machine learning techniques, data preprocessing, and model interpretation through hands-on involvement in the project.
	Rationale	Traditional customer retention strategies in the automotive sector rely heavily on intuition, manual segmentation, and basic statistical methods, which are often inefficient and unable to fully utilize available data. There is a need for data-driven solutions to identify customers most likely to repurchase, optimize marketing spend, and improve sales efficiency. The program leverages machine learning and predictive analytics to address class imbalance and large-scale datasets, providing actionable insights that can enhance customer engagement and operational decision-making. This initiative also serves as a platform for applied learning, industry collaboration, and innovation in educational and research practices at WUB.
Subject (Leader)	Initiator(s)	Officine Alfieri Maserati SA
	Champion(s)	HOSSAIN, Md Tanzim
	Major team member(s)	IRAIIVAN, Ezilaan, GARCIA Y GARCIA, Jamie, BIN SALEH, Shaqran, KOMBAN LAWRENCE, Jasmine Rose, SO, Anthony, IQBAL, Muhammad
Environment	Nature/Society	This project impacts society by promoting more efficient use of resources in vehicle marketing and sales, leading to reduced waste from overproduction and excessive promotional efforts.
	Industry/Market	The project directly relates to the automotive sales and service industry, particularly car dealerships and resellers seeking to enhance customer retention. It supports the shift toward data-driven operations by using machine learning to predict repurchase behavior, allowing businesses to make informed decisions in marketing, sales, and inventory management.
	Citizen/Government	The project may align with digital transformation initiatives promoted by governmental bodies and industry regulators. It could also be of interest to

		academic institutions and research organizations focused on applied machine learning, business analytics, and customer relationship management within the automotive sector.
Resources	Human resources	The project required expertise in data science, particularly in supervised machine learning and model evaluation techniques. Additionally, professionals with domain knowledge in automotive sales and customer behavior were essential to guide feature selection and ensure the practical relevance of the models. Support from business analysts and stakeholder liaisons was also important for aligning the technical work with real-world business needs.
	Financial resources	Funding was necessary for data processing, model development, and access to required computational infrastructure. Financial support also contributed to covering personnel costs, including data scientists, analysts, and project coordinators. As this is a demo-phase project partially funded by Maserati, additional financial resources will be required for scaling and deployment in production environments.
	Technological resources	The project relied on access to computing environments capable of handling large-scale datasets and performing complex machine learning computations. Software tools such as Python, Scikit-learn, and Jupyter Notebooks were used for data analysis and model building. Secure data storage and version control systems were also needed to manage dataset integrity and experiment reproducibility.
Mechanism	Strategy (Weight/Sequence)	Program Focus: Predicting customer repurchase behavior in the automotive resale sector using supervised machine learning. Key Techniques: Decision Trees, Random Forests, SVM, Extra Trees. Prioritized Steps (Sequence & Weight): 1. Data Cleaning & Exploration (High Weight, First Step): Handle missing values, understand class distribution, and assess feature relevance. 2. Feature Selection & Encoding (High Weight, Second Step): Identify important variables, encode categorical data for model readiness. 3. Model Implementation & Training (High Weight, Third Step): Experiment with multiple algorithms to determine predictive performance. 4. Evaluation & Refinement (Medium-High Weight, Fourth Step): Use ROC AUC to compare models and tune hyperparameters via Grid Search. 5. Model Selection & Reporting (Medium Weight, Final Step): Choose the best-performing model (Decision Tree in demo), analyze insights for business applications.
	Organization	The university's organizational structure aligns moderately well with the program's strategies. Strengths: Provides access to computing resources, faculty expertise in data science, and structured research support. Limitations: Collaboration with external stakeholders (automotive company) may require formal partnerships and approvals, which can introduce delays.
	Culture	The university culture generally supports data-driven research and machine learning experimentation. Supportive Factors: Encourages evidence-based approaches, innovative projects, and access to datasets.

		Potential Hindrances: Traditional administrative processes may slow integration with industry data or restrict rapid iterative experimentation.
Doing		
Launch date		September 2024
Responsible organization		Officine Alfieri Maserati SA
Program content and process		The program focuses on predicting customer repurchase behavior in the automotive resale sector using supervised machine learning techniques. It utilizes a demo dataset of over 130,000 anonymized customer records containing variables related to car models, service history, vehicle age, mileage, and dealer interactions. The implementation process begins with data cleaning and exploration to handle missing values and examine class distribution, addressing the significant class imbalance where only 2.68% of customers had repurchased. Feature selection and encoding are then performed to prepare the dataset for machine learning. Multiple classification algorithms, including Decision Trees, Random Forests, Support Vector Machines (SVM), and Extra Trees, are implemented to compare predictive performance. Evaluation focuses on ROC AUC, balancing sensitivity and specificity in the presence of data imbalance. Hyperparameter tuning techniques such as Grid Search are applied to optimize model performance. The Decision Tree model with default parameters was selected for the demo phase due to its strong predictive performance (ROC AUC of 0.84), computational efficiency, and interpretability. Insights derived from the model inform targeted marketing, customer relationship management, and inventory planning. The project also emphasizes structured and systematic data preprocessing, including handling missing values, encoding categorical variables, and splitting data into training, testing, and validation sets. The process ensures replicability and provides a foundation for future expansion, including the integration of ensemble stacking methods, Explainable AI tools (SHAP, LIME), and potential deep learning approaches. Overall, the program combines advanced machine learning techniques with practical business applications, creating a scalable, data-driven framework to optimize customer retention and sales efficiency while providing experiential learning opportunities for students and faculty.
Key highlights of the content/process		Content Highlights: 1. Focus on predicting customer repurchase behavior using large-scale, anonymized datasets. 2. Use of multiple machine learning models (Decision Tree, Random Forest, SVM, Extra Trees). 3. Structured preprocessing to address missing data and class imbalance. Process Highlights: 1. Systematic model evaluation using ROC AUC and hyperparameter tuning. 2. Pilot implementation for actionable insights in marketing and sales. 3. Scalable framework for future integration with live operational datasets.

Differences from traditional approaches	Traditional methods rely on manual segmentation and intuition; this program uses data-driven predictive modeling. Unlike conventional single-rule heuristics, multiple algorithms were compared for optimized accuracy and efficiency. The project systematically utilizes large datasets, advanced preprocessing, and evaluation metrics, capturing complex feature interactions often overlooked in traditional methods.
Progress as of today	• Data cleaning, feature selection, and encoding completed. • Multiple classification models implemented and evaluated. • Decision Tree model selected as the optimal demo model. • Program is ongoing; live dataset integration and production deployment remain future steps.
Problems in implementation	• Limited diversity in the demo dataset; may not represent real-world scenarios. • Severe class imbalance (only 2.68% repurchases). • Computational resource requirements for training and hyperparameter tuning. • Challenges in integrating models into existing CRM workflows.
Approaches to solve the problems	• Engaging data scientists, domain experts, and software engineers for technical and operational guidance. • Applying techniques like SMOTE and under sampling to address class imbalance. • Pilot deployment on small segments to validate performance and build stakeholder trust. • Ensuring compliance with data protection laws and secure anonymization to enable richer data collection.
Completion date, if completed	Ongoing
Seeing	
Impacts on students	The project may inspire students to explore advanced machine learning in marketing analytics like automotive retail, e-commerce, and customer experience management.
Impacts on professors	Professors may integrate this project as a case study into data science, marketing analytics, or applied AI or ML courses.
Impacts on university administration	Faculty and the university can gain recognition for pioneering the use of predictive analytics and ensemble machine learning in business and automotive contexts.
Responses from industry/market	Automotive and marketing firms are likely to show interest in the project's predictive models for their potential to improve customer retention and streamline sales strategies. This could lead to partnerships, pilot testing, or commercialization opportunities.
Responses from citizen/government	N/A
Measurable output (revenues)	• The key measurable output for the demo initial study is the ROC AUC score, with the selected Decision Tree model achieving a high ROC AUC of 0.84, indicating strong ability to distinguish between repurchasing and non-repurchasing customers. • The project monitors precision, recall, and F1-score for the minority class (repurchase), ensuring the model's effectiveness beyond basic accuracy metrics. • Improvement in identifying high-value customer segments can be

	tracked through changes in targeted marketing conversion rates. • Measurable improvements in campaign ROI, customer retention rate, and inventory alignment, once deployed, will serve as business performance indicators.
Measurable input (expenses)	Confidential data
Cost-benefit analysis for effectiveness	Costs: • Significant time and expertise were invested in developing and evaluating multiple machine learning models, especially to address data imbalance and tune hyperparameters. • Resources were needed to process large datasets, run cross-validation, and conduct hyperparameter tuning, often requiring extended cloud computing usage. Benefits: • The model enables more precise targeting of customers likely to repurchase, potentially increasing retention rates and reducing marketing cost. • By identifying high-probability repurchasers, the business can allocate marketing and sales resources more effectively. • Once validated and deployed, the model offers scalability with minimal additional cost, resulting in long-term savings through improved marketing ROI and reduced churn.
Future Planning	
Where does the project go from here?	Transitioning from the current demo dataset to a live production dataset sourced directly from client systems, with appropriate permissions and anonymization protocols, will enhance the model's relevance and reliability. Incorporating additional variables such as customer financial data, online interaction history, or post-sale engagement metrics could offer deeper insights into repurchase behavior. Implementing sophisticated approaches like SMOTE-ENN or cost-sensitive learning to better address class imbalance and enhance minority class detection. Deploying the model in a limited business environment to monitor its real-time performance, marketing impact, and operational integration challenges. Integrating the model output with customer relationship management systems to automate targeted campaigns and follow-ups. Deep Learning Exploration Investigating the use of neural network models, including feed-forward and recurrent architectures, to capture nonlinear patterns and temporal customer behaviors. Ensemble Machine Learning Expansion: Future phases will explore ensemble stacking methods combining predictions from multiple base learners (e.g., Decision Trees, Random Forests, Extra Trees, SVM) for improved generalization. Explainable AI (XAI) Incorporating tools like SHAP or LIME to provide transparent reasoning behind predictions, which is crucial for stakeholder trust and regulatory compliance.
Addendum	

Exhibits, pictures, diagrams, etc.

Table 1

Table showing data split percentage

Serial No	Split Section	Percentage of Data
1	Training Data	70%
2	Testing Data	15%
3	Validation Data	15%

Table 2

Summary of Classifier Models Experimented in this Project

Serial No	Name of the Classifier Model	Hyperparameters Setting
1	Support Vector Machine (SVM)	Default
2	Decision Tree	Default
3		min_samples_split = 5
4		min_samples_split = 10
5		min_samples_split = 20
6	Random Forests	Default
7		n_estimators = 50
8		n_estimators = 200
9		max_depth = 14, n_estimators = 50
10		Using Grid Search (max_depth = 25, n_estimators = 60, min_samples_leaf = 8)
11	Extra Trees	Default

Figure 1

Summary of Models' Performances

Performance of Models evaluated by the ROC AUC scores on the training, validation, and test datasets

Serial No	Name of the Classifier Model	Hyperparameters Setting	ROC AUC Scores		
			Training Set	Validation Set	Test Set
1	Support Vector Machine (SVM)	Default	0.65	0.633	0.64
2	Decision Tree	Default	0.982	0.833	0.84
3		min_samples_split = 5	0.924	0.826	0.84
4		min_samples_split = 10	0.893	0.824	0.83
5		min_samples_split = 20	0.879	0.824	0.84
6	Random Forests	Default	0.983	0.804	0.82
7		n_estimators = 50	0.983	0.791	0.82
8		n_estimators = 200	0.983	0.804	0.82
9		max_depth = 14, n_estimators = 50	0.836	0.756	0.77
10		Using Grid Search (max_depth = 25, n_estimators = 60, min_samples_leaf = 8)	0.761	0.724	0.75
11	Extra Trees	Default	0.982	0.769	0.78

Reports, mimeos, monographs, books, etc.	Due to contractual agreement with the sponsoring organization, detailed project documentation such as reports, monographs, or internal publications cannot be shared without authorized permission from the stakeholders.
Others which may help explain the program (including website links)	This project is subject to confidentiality conditions under an industry-academic collaboration. As a result, additional supporting materials, including digital content or internal resources, cannot be disclosed publicly without prior authorization from the sponsoring entity.