FINAL REPORT

*Submitted b*y

SAADANA SHYAMSUNDER PUJA MADHAV

SIREESHA

PALLAVI KONDRA

KAVYA KUMARY JAYANTHI NAIR

*In partial fulfillment of the credit requirements in*

**DSCI 5350**

*Under the guidance of*

**Dr. Scott Hamilton**



**DEPARTMENT OF ITDS**

G BRINT RYAN COLLEGE OF BUSINESS

**DECEMBER 2024**

# Table Of Contents

[Table Of Contents 1](#_bookmark0)

1. [Executive Summary 3](#_bookmark1)
2. [Objective 4](#_bookmark2)
   1. [Strategic Importance of Solving the Problem **4**](#_bookmark3)
3. [Data Understanding and Data Overview 5](#_bookmark4)
   1. [Introduction to the Dataset 5](#_bookmark5)
   2. [Data Exploration 6](#_bookmark6)
4. [Data Preparation 10](#_bookmark11)
   1. [Accessing and Storing Data 10](#_bookmark12)
      1. [Steps in Data Preparation 10](#_bookmark13)
   2. [Big Data Tools and Ecosystem 11](#_bookmark14)
5. [Statistical Analysis and Modeling 12](#_bookmark15)
   1. [Overview of Modeling Process 12](#_bookmark16)
      1. [Class Imbalance Handling with SMOTE 12](#_bookmark17)
      2. [Pipeline Construction 13](#_bookmark18)
      3. [Model Training and Evaluation: 13](#_bookmark19)
   2. [Initial Model Selection 13](#_bookmark20)
   3. [XGBoost's Performance Improvements 14](#_bookmark21)
   4. [Feature Importance of XGBoost 14](#_bookmark22)
   5. [Model Evaluation and Metrics 16](#_bookmark25)
   6. [Cluster Analysis - Comprehensive Technique for Customer Segmentation 18](#_bookmark28)
      1. [Purpose of Cluster Analysis 18](#_bookmark29)
      2. [Implementation Steps 18](#_bookmark30)
      3. [Importance of Silhouette Score 19](#_bookmark31)
      4. [PCA Scatter Plot 19](#_bookmark32)
      5. [Cluster Characteristics Extraction 20](#_bookmark34)
6. [Model Interpretation 22](#_bookmark36)

[7 Business Implications and Recommendations 25](#_bookmark37)

1. [Future Scope of Improvement 28](#_bookmark38)
2. [Conclusion 30](#_bookmark39)

[11 Additional References 31](#_bookmark40)

**Table of Figures**

[**Fig. 3.1** Target Variable Distribution 6](#_bookmark7)

[**Fig. 3.2** Age vs. Annual Premium 7](#_bookmark8)

[**Fig. 3.3** Distribution of Numerical Features 8](#_bookmark9)

[**Fig. 3.4** Correlation Matrix 9](#_bookmark10)

[**Fig 5.1** Feature Importance output 15](#_bookmark23)

[**Fig 5.2** Top features 15](#_bookmark24)

[**Fig 5.3** Model Output Screenshot 17](#_bookmark26)

[**Fig 5.4** AUC-ROC curve 17](#_bookmark27)

[**Fig 5.5** PCA – based Cluster Scatterplot 19](#_bookmark33)

[**Fig 5.6** Cluster Silhouette Score Output 20](#_bookmark35)

# Executive Summary

The project aimed to address the critical business challenge of identifying potential customers for vehicle insurance cross-selling within an existing base of health insurance clients. Leveraging big data analytics, the objective was to improve marketing efficiency and maximize conversions while minimizing costs.

The dataset contained over 300,000 records with features such as demographics, policy details, and vehicle attributes. Data preparation included handling class imbalance through SMOTE, encoding categorical variables, and scaling continuous features. Advanced visualization libraries like Plotly and Bokeh were used to explore data distributions and correlations.

Key methodologies included predictive modeling using machine learning algorithms (Logistic Regression, Decision Trees, and XGBoost) and cluster analysis through K-Means. XGBoost emerged as the most effective model, achieving an AUC of 83.63%, with significant features such as Vehicle Damage, Vehicle Age, and Previously Insured status influencing predictions. Cluster analysis revealed five distinct customer segments, enabling tailored marketing strategies.

The findings highlighted actionable business implications:

* 1. Targeted Campaigns: Focus on customers with high vehicle damage likelihood and older vehicle age.
  2. Optimized Resources: Prioritize high-potential leads based on predictive insights.
  3. Cross-Sell Opportunities: Address gaps in insurance coverage for partially insured customers.

Future enhancements include incorporating additional features like income and credit scores, statistical analysis for feature selection, and hyperparameter tuning for model refinement. The integration of predictive analytics into marketing strategies demonstrated significant potential for revenue growth, improved customer satisfaction, and efficient resource utilization. This project underscores the value of combining machine learning and data-driven decision-making to tackle real-world business challenges in the insurance sector.

**1.1 Project Motivation/Background**

The insurance industry faces an ongoing challenge in identifying and converting potential customers for additional policy offerings, such as vehicle insurance. While cross-selling remains a key strategy to enhance customer lifetime value, traditional marketing approaches often lack precision, leading to inefficient resource allocation and low conversion rates. This project was chosen to address the need for a data-driven approach to optimize cross-selling efforts. By leveraging customer data from health insurance policies, the goal was to predict interest in vehicle insurance, identify influential factors driving customer decisions, and segment the customer base for targeted marketing campaigns.

The importance of this project lies in its potential to revolutionize marketing efficiency and customer engagement. Predictive modeling allows for precise targeting of high probability leads, ensuring better resource utilization. Additionally, segmentation insights enable personalized strategies tailored to unique customer profiles, fostering higher satisfaction and retention rates.

# Objective

## 2.1 Strategic Importance of Solving the Problem

The main objective of this project is to utilize big data analytics and machine learning techniques to identify potential customers within a health insurance customer base who are most likely to purchase vehicle insurance. By leveraging advanced predictive models and customer segmentation, the project seeks to enhance marketing efficiency, optimize resource allocation, and maximize conversion rates. Additionally, the project aims to uncover key factors influencing customer decisions, enabling data-driven strategies for personalized and targeted marketing campaigns.

This objective is aligned with the broader goal of improving cross-selling effectiveness in the insurance sector while addressing challenges such as imbalanced data, limited customer insights, and inefficient resource use.

**Problem Statement**

The insurance industry often faces difficulties in executing successful cross-selling strategies due to the lack of precise customer targeting and understanding of purchase behaviors. These inefficiencies result in wasted marketing resources, low conversion rates, and missed revenue opportunities.

# Data Understanding and Data Overview

## Introduction to the Dataset

Kaggle Link to Dataset – [Click Here!](https://www.kaggle.com/datasets/anmolkumar/health-insurance-cross-sell-prediction) Colab Notebook Link – [Click here!](https://colab.research.google.com/drive/1tY1FM47ql-YD78_5Wna_mZ2p4g7xca9S?usp=sharing)

The dataset is designed to help an insurance company predict whether customers who currently have health insurance policies will also be interested in purchasing vehicle insurance. It includes detailed customer information such as demographics, vehicle attributes, and policy-related details.

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Description** | **Details/Values** |
| id | Unique identifier for each customer. | - |
| Gender | Gender of the customer. | Male or Female |
| Age | Age of the customer in years, representing their demographic profile. | - |
| Driving\_License | Indicates if the customer has a valid driving license. | 1: Valid license, 0: No valid license |
| Region\_Code | A unique code representing the geographic region of the customer. | Help segment customers geographically for targeted marketing |
| Previously\_Insured | Indicates if the customer already holds a vehicle insurance policy. | 1: Has insurance, 0: No insurance |
| Vehicle\_Age | Age of the customer's vehicle. | < 1 Year: New, 1-2 Year: Moderately used, > 2 Years: Older |
| Vehicle\_Damage | It indicates if the customer has reported vehicle damage in the past. | 1: Reported damage, 0: No reported damage |
| Annual\_Premium | Yearly premium amount paid by the customer. | Indicates financial capacity and prior spending on insurance |
| Policy\_Sales\_Channel | Anonymized code indicating the channel used to contact the customer. | e.g., agent, email, phone, in-person |
| Vintage | Number of days the customer has been associated with the insurance company. | Reflects customer loyalty and tenure |
| Response (Target) | Indicates customer interest in purchasing vehicle insurance. | 1: Interested, 0: Not interested |

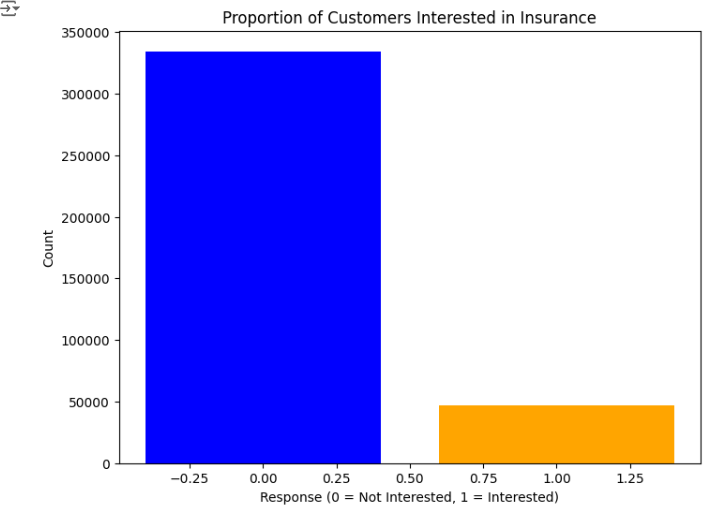
## Data Exploration

Data exploration is a critical step in any data analytics project, providing valuable insights into the dataset's structure, distributions, relationships, and potential challenges. This phase focuses on understanding the data at a deeper level to guide preprocessing and model development. The purpose of this is to:

* + 1. Identify patterns and trends within the data.
    2. Detect anomalies, outliers, and missing values.
    3. Assess the relationships between features and the target variable.
    4. Evaluate feature distributions for statistical properties like skewness and normality.
    5. Gain actionable insights to guide feature engineering and modeling decisions.

**Target Variable Distribution**:

This visualization depicts the distribution of customer interest in vehicle insurance using a bar chart. The chart categorizes customers into two groups: those interested in purchasing vehicle insurance (Response = 1) and those not interested (Response = 0). The analysis is critical for understanding the overall inclination of the customer base, enabling targeted marketing and resource allocation. A significant majority of customers fall into the "Not Interested" category (Response = 0), indicating that most policyholders do not express immediate interest in vehicle insurance. This highlights the challenge of converting uninterested customers into leads. A smaller, but crucial, proportion of customers are in the "Interested" category (Response = 1). These customers represent potential opportunities for the company to generate revenue through targeted campaigns.

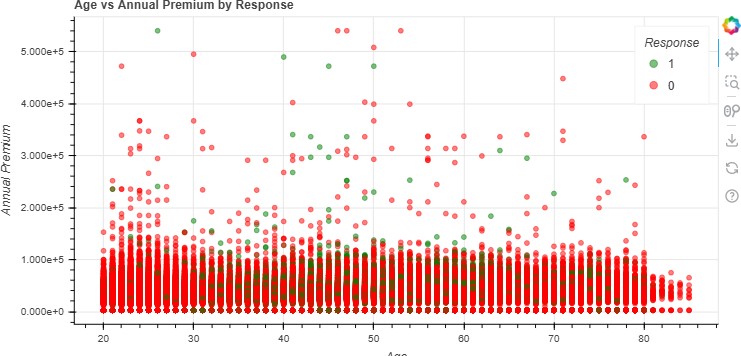


***Fig. 3.1*** *Target Variable Distribution*

**Age vs. Annual Premium by Response:**

This visualization uses a scatter plot to illustrate the relationship between customer age, annual premium, and their response to vehicle insurance interest. By analyzing this chart, we can identify potential patterns and trends among customer demographics, premium values, and interest levels. The plot was created using

**Bokeh**, an interactive visualization library, which allows dynamic representation of data. Data was converted from a Spark DataFrame to Pandas for compatibility with Bokeh.

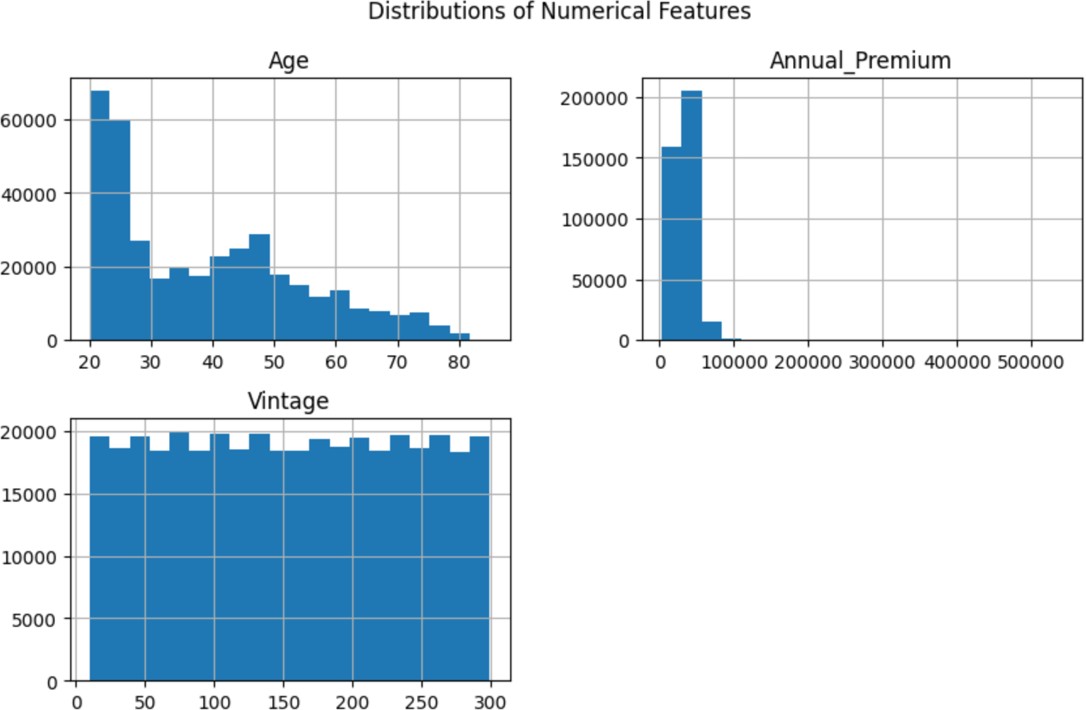


***Fig. 3.2*** *Age vs. Annual Premium*

Green (Response = 1): Indicates customers interested in vehicle insurance. Red (Response = 0): Represents customers not interested in vehicle insurance. High annual premiums do not necessarily correlate with customer interest. Customers with a wide range of premium values appear in both categories. Customers of all ages exhibit a mix of interest levels, with no single age group showing a definitive trend for being more interested or disinterested.

**Distribution of Numerical Features**

This provides an analysis of the distribution of three key numerical features in the dataset: Age, Annual Premium, and Vintage. The histograms illustrate the range and frequency of these features, offering valuable insights into customer demographics, premium payment patterns, and the length of customer association with the company.



***Fig. 3.3*** *Distribution of Numerical Features*

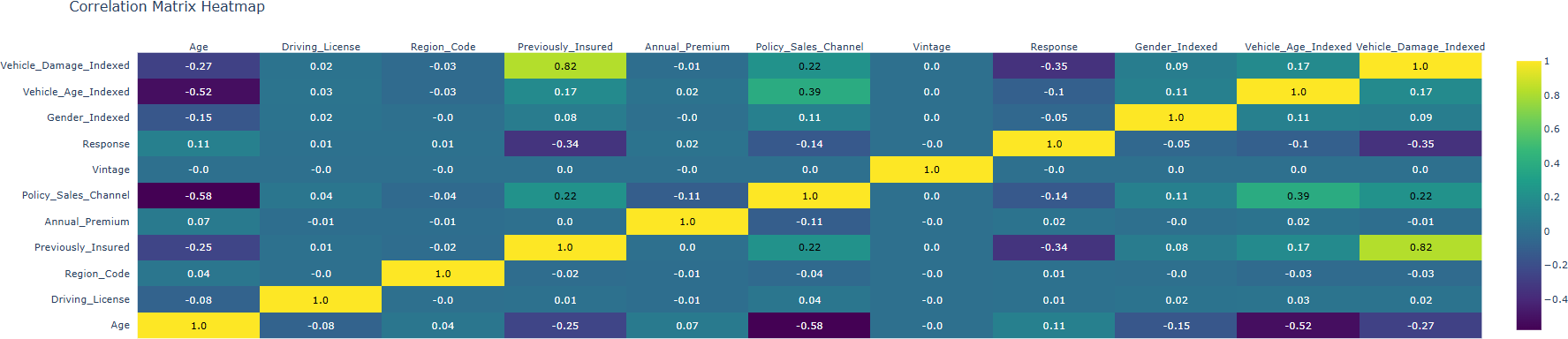
The distribution of customer age exhibits a right-skewed pattern, with most customers concentrated in the **20–40** age range. The frequency decreases significantly for customers aged above 50, indicating a smaller representation of older individuals. The predominance of younger customers highlights a strategic focus on this demographic.

The annual premium values are heavily right skewed, with most customers paying premiums in the range of **10,000 to 50,000**. A small number of customers fall into the high-premium category, representing outliers who may hold high-value policies. The concentration of premiums in the lower range suggests that affordable plans are a key driver for customer acquisition.

The distribution of the "Vintage" feature, representing the number of days a customer has been associated with the company, is uniform across its range (0 to 300 days). This indicates a balanced mix of both new and long-term customers.

**Correlation Matrix Heatmap Analysis**

The correlation matrix heatmap provides a detailed view of the relationships between various numerical features in the dataset, offering valuable insights into feature relevance and interactions. Below is a professional breakdown of its purpose, methodology, and findings.



***Fig. 3.4*** *Correlation Matrix*

The interactive heatmap was generated using Plotly for a visually engaging representation of the correlation matrix. Numerical annotations within each cell display the exact correlation coefficient, rounded to two decimal places for clarity. Color intensity represents the strength of correlations:

* Bright/yellow: Strong positive correlations.
* Darker/purple: Strong negative correlations.
* Neutral colors: Weak or no correlation.

*.* ***Key Insights from the Heatmap***

1. Strong Correlations with the Target Variable (Response):
   * *Vehicle\_Damage\_Indexed* (0.35): A moderate positive correlation suggests that customers with prior vehicle damage are more likely to purchase vehicle insurance.
   * *Previously\_Insured* (-0.34): A moderate negative correlation indicates that customers already holding vehicle insurance are less likely to show interest in a new policy.
2. Feature Interactions:
   * *Vehicle\_Age\_Indexed* and *Vehicle\_Damage\_Indexed* (0.17): A weak positive correlation highlights a slight tendency for older vehicles to have experienced damage.
   * *Vehicle\_Age\_Indexed* and *Policy\_Sales\_Channel* (0.39): A moderate correlation suggests that specific sales channels effectively target customers with older vehicles.
   * *Age and Policy\_Sales\_Channel* (-0.58): A strong negative correlation indicates that older customers are less likely to be approached via certain sales channels.
3. Weak Correlations:
   * Features like *Region\_Code* and *Driving\_License* exhibit weak correlations with both the target variable (Response) and other features, indicating limited predictive value.
4. Independent Features:
   * Features like *Annual\_Premium* and *Vintage* show weak correlations with most variables, suggesting they behave independently within the dataset.

This heatmap analysis is pivotal for guiding feature selection and model optimization, enabling the identification of impactful variables while minimizing redundancy and improving predictive accuracy.

# Data Preparation

Preparing the dataset for predictive modeling involved several critical steps to ensure data quality, handle imbalances, and optimize the data for machine learning algorithms. Below is an overview of the data preparation process and how the big data ecosystem and tools were utilized.

* 1. **Accessing and Storing Data Data Storage**:

The dataset was stored in Google Drive to ensure easy access and cloud-based collaboration. Google Drive

served as a reliable and centralized storage location for seamless sharing and integration with data processing platforms.

**Data Access**:

The data was accessed through Google Colab, a cloud-based platform that supports PySpark and Python, enabling distributed and scalable data processing. Using Google Colab allowed us to directly mount Google Drive and access the dataset in a secure and efficient manner.

### Steps in Data Preparation

* + - 1. **Loading the Data**:
* The dataset was loaded into a PySpark DataFrame for distributed processing.
* PySpark enabled handling large datasets efficiently, leveraging its scalable architecture.
  + - 1. **Data Cleaning**:
* A thorough check revealed no missing values in the dataset, ensuring data completeness.
* Columns like id (unique identifier) were removed as they do not contribute to the predictive analysis.
  + - 1. **Handling Categorical Variables**:
* Categorical features such as *Gender, Vehicle\_Age, and Vehicle\_Damage* were encoded using StringIndexer to convert them into numerical format suitable for machine learning algorithms.
* Anonymized categorical variables like *Region\_Code* and *Policy\_Sales\_Channel* were retained in their original format for model input.
  + - 1. **Scaling Numerical Features**:
* Continuous variables (*Age, Annual\_Premium, Vintage*) were scaled using StandardScaler to standardize their ranges and prevent dominance of larger-scale features over smaller ones.
* This ensured that machine learning models, especially those sensitive to feature magnitudes, performed optimally.
  + - 1. **Handling Class Imbalance**:
* The target variable Response was highly imbalanced, with most customers not interested in vehicle insurance.
* To address this *SMOTE (Synthetic Minority Oversampling Technique)* was applied to the training dataset. SMOTE created synthetic samples of the minority class (Response = 1), balancing the data and ensuring model training was not biased toward the majority class.
  + - 1. **Outlier Detection and Treatment**:
* Variables like *Annual\_Premium* exhibited skewed distributions with potential outliers in high ranges.
* Outliers were retained after analysis, as they represented genuine customer behavior (e.g., high premium payers).
* These outliers were mitigated during scaling to ensure they did not adversely affect the model.
  + - 1. **Normalization and Non-Normal Distributions**:
* Non-normal distributions, particularly in features like *Annual\_Premium*, were addressed through scaling and model-specific adjustments, ensuring that the data was uniformly distributed for machine learning algorithms.
* Techniques such as clustering and gradient boosting handled non-linear relationships effectively without requiring explicit transformation of non-normal data.

## Big Data Tools and Ecosystem

1. PySpark:
   * Used for distributed data processing, enabling fast computation on large datasets.
   * Functions like VectorAssembler and StringIndexer streamlined feature preparation.
2. Google Colab:
   * Provided an interactive platform for PySpark processing with seamless integration of libraries like Matplotlib and Plotly for visualization.
   * Allowed access to Google Drive for data storage and retrieval.
3. Visualization Libraries:
   * Libraries like Plotly and Bokeh were used to explore data distributions and relationships visually, enabling deeper understanding during the preparation phase.

# Statistical Analysis and Modeling

This section provides an in-depth explanation of the statistical techniques and machine learning methodologies used to predict customer propensity for purchasing additional insurance. A careful model selection process evaluated diverse algorithms to identify the best-performing model for the task. In addition to predictive modeling, clustering techniques were applied to segment customers into actionable groups, enabling targeted marketing strategies. Finally, rigorous evaluation metrics were employed to assess the performance of the models and validate their applicability.

## Overview of Modeling Process

The modeling process was methodically designed to handle the challenges of imbalanced data, diverse feature types, and scalability, ensuring robust predictions. The steps involved:

### Class Imbalance Handling with SMOTE

The Response variable, which indicates whether a customer is interested in vehicle insurance, was highly imbalanced. Most customers were not interested (Response = 0). To address this, we applied **SMOTE (Synthetic Minority Oversampling Technique)**. This technique generates synthetic samples for the minority class, balancing the dataset without introducing bias.

**Why Use SMOTE?**

In our dataset, the target variable Response was highly imbalanced, with a majority of customers (around 87%) not interested in vehicle insurance (Response = 0), and only a small minority (around 13%) interested (Response = 1). This imbalance posed the following challenges:

* **Bias in Model Training**: Machine learning models tend to favor the majority class, leading to poor predictive performance for the minority class.
* **Reduced Recall for Minority Class**: The model could miss out on correctly identifying customers who are likely to be interested in vehicle insurance.

SMOTE was applied to oversample the minority class by generating synthetic data points, thus balancing the target variable's class distribution.

* + - 1. **Class Distribution Before SMOTE**:
         * Majority Class (Response = 0): ~267,471 samples
         * Minority Class (Response = 1): ~37,267 samples
         * This imbalance resulted in the model favoring predictions for Response = 0.
      2. **Application of SMOTE**:
         * Using the SMOTE algorithm, the minority class was oversampled to match the majority class, creating a balanced dataset.
      3. **Class Distribution After SMOTE**:
         * Majority Class (Response = 0): ~267,471 samples
         * Minority Class (Response = 1): ~267,471 samples
         * This balance allowed the model to treat both classes equally during training.

### Pipeline Construction:

To streamline the modeling process, we created a PySpark pipeline that combined all preprocessing steps and model training into a single workflow. The pipeline included:

* ***Data Transformation***: Scaling, encoding, and SMOTE were integrated as pipeline stages.
* ***Model Training***: Each model (Logistic Regression, Decision Tree, XGBoost) was trained using transformed data.
* ***Validation:*** The pipeline ensured consistent preprocessing across training and testing datasets, reducing the risk of data leakage.

This modular approach allowed us to seamlessly test multiple algorithms and ensure reproducibility.

### Model Training and Evaluation:

Three machine learning models were trained and evaluated: Logistic Regression, Decision Tree Classifier, and Gradient Boosted Trees (XGBoost). The evaluation metrics included Accuracy, Precision, Recall, F1-Score, and Area Under the Curve (AUC), with AUC selected as the primary metric due to its robustness for imbalanced datasets.

## Initial Model Selection

The following models were selected based on their capabilities to address the problem’s requirements:

1. **Logistic Regression**:
   * **Purpose**: Established a baseline for performance comparison.
   * **Strengths**: Simple, interpretable, and computationally efficient.
   * **Limitations**: Inability to capture non-linear relationships.
2. **Decision Tree Classifier**:
   * **Purpose**: Captured non-linear patterns in the data and provided feature importance insights.
   * **Strengths**: High interpretability and ability to handle non-linear decision boundaries.
   * **Limitations**: Prone to overfitting, especially on imbalanced data.
3. **Gradient Boosted Trees (XGBoost)**:
   * **Purpose**: Used iterative boosting to improve prediction accuracy and handle imbalanced datasets.
   * **Strengths**: Superior performance on structured data, robust to noise and overfitting, and capable of capturing complex patterns.
   * **Performance**: XGBoost achieved the highest AUC score of 83.63%, outperforming other models.

## XGBoost's Performance Improvements

XGBoost (Extreme Gradient Boosting) was a pivotal algorithm in our project, delivering the best performance among the models we evaluated for predicting customer interest in vehicle insurance. Its advanced tree-based boosting mechanism and ability to handle complex patterns in the data made it a natural choice for this classification task. Here’s a detailed breakdown of how XGBoost excelled and what improvements it brought to the project.

**Why XGBoost?**

XGBoost was chosen for its:

1. **Efficiency**: XGBoost is optimized for speed and performance, using techniques like parallelization and cache optimization.
2. **Robustness**: It handles missing data and outliers effectively, making it ideal for datasets with some inconsistencies or varying patterns.
3. **Accuracy**: Its iterative boosting mechanism allows it to minimize errors progressively, resulting in high predictive accuracy.
4. **Feature Importance**: XGBoost provides insights into feature importance, making it easier to interpret the drivers of customer behavior.

**Performance Enhancements by XGBoost**

1. **Handling Complex Data Relationships**:
   * XGBoost's gradient boosting mechanism allowed it to capture non-linear relationships in the data better than Logistic Regression or Decision Tree models.
   * It iteratively corrected errors made by previous trees, creating a strong final model.
2. **Feature Importance Analysis**:
   * XGBoost provided a ranked list of features by importance, allowing us to understand which variables significantly influenced the predictions.
   * Key features like *Vehicle\_Damage\_Indexed*, *Vehicle\_Age\_Indexed*, and *Previously\_Insured* were identified as major drivers, enabling the company to tailor its marketing strategies effectively.
3. **Regularization for Generalization**:
   * XGBoost uses L1 (Lasso) and L2 (Ridge) regularization to prevent overfitting. This ensured that the model performed well on both training and test datasets.
4. **Handling Class Imbalance**:
   * After applying SMOTE to address class imbalance, XGBoost leveraged the balanced data effectively, improving its ability to predict the minority class (Response = 1) while maintaining high precision and recall.

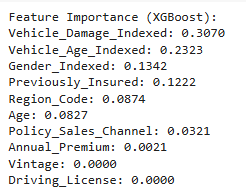
## Feature Importance of XGBoost

Feature importance analysis in XGBoost is a critical step that provides insights into which variables have the most significant impact on the model's predictions. This helps both with understanding the model's decision-making process and in identifying key business drivers.

**What is Feature Importance?**

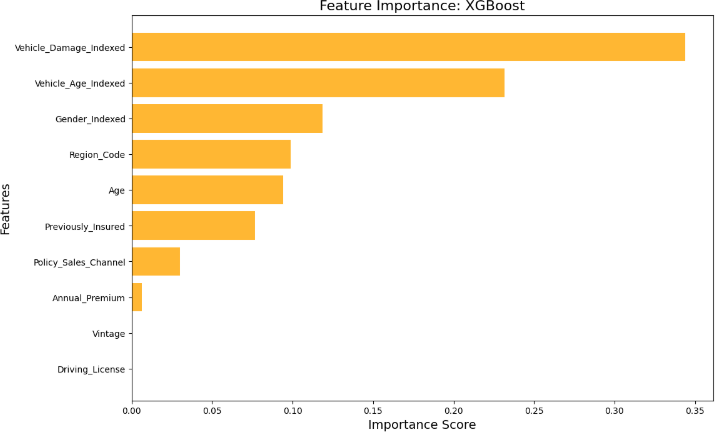
Feature importance refers to the contribution of each feature to the prediction of the target variable in the XGBoost model. It quantifies how much each feature contributes to reducing the error or improving the model's performance. In XGBoost, feature importance is derived from:

1. **Gain**: The improvement in accuracy brought by a feature to the branches it splits.
2. **Frequency**: The number of times a feature is used in the decision tree splits.
3. **Cover**: The proportion of samples helps split correctly.



***Fig 5.1*** *Feature Importance output*

The feature importance was visualized using a bar chart, where each feature’s relative importance was displayed. The chart clearly highlighted the dominance of *Vehicle\_Damage\_Indexed* and *Vehicle\_Age\_Indexed*, followed by *Gender\_Indexed* and *Previously\_Insured*.



***Fig 5.2*** *Top features*

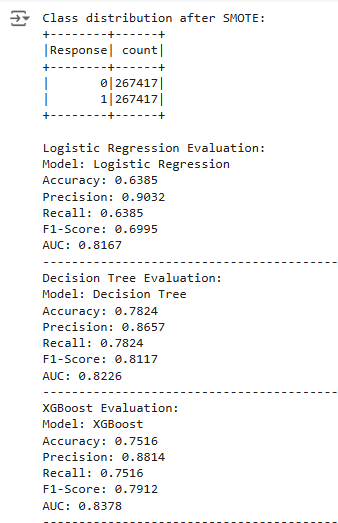
## Model Evaluation and Metrics

The performance of the models was assessed using multiple evaluation metrics to ensure a comprehensive understanding of their predictive capabilities. These metrics were carefully chosen to address the specific challenges of the dataset, such as class imbalance and the need to balance false positives and false negatives in predictions.

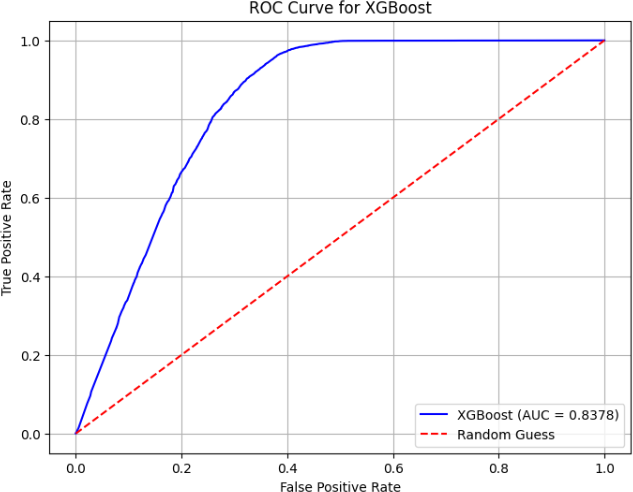
* + 1. ***Why AUC Was the Primary Metric***
       1. **Imbalanced Dataset**:
          - The target variable exhibited significant class imbalance, with most customers not purchasing additional insurance. Metrics like accuracy could be misleading, while AUC effectively accounted for the imbalance.
       2. **Threshold Independence**:
          - Unlike accuracy, AUC evaluates performance across all possible thresholds, providing a more holistic assessment of the model.
       3. **Business Implications**:
          - By focusing on AUC, the model prioritized its ability to distinguish between potential buyers and non-buyers, ensuring that marketing efforts could be effectively targeted.
    2. ***Evaluation Summary***

|  |  |  |  |
| --- | --- | --- | --- |
| **Metric** | **Logistic Regression** | **Decision Tree** | **XGBoost** |
| **Accuracy** | 0.6385 | 0.7896 | 0.7535 |
| **Precision** | 0.9032 | 0.8570 | 0.8787 |
| **Recall** | 0.6385 | 0.7896 | 0.7535 |
| **F1-Score** | 0.6995 | 0.8149 | 0.7924 |
| **AUC** | 0.8167 | 0.8222 | 0.8363 |

XGBoost’s balance of high precision and recall, coupled with its superior AUC score (0.8363) validated its selection as the best-performing model.



***Fig 5.3*** *Model Output Screenshot*



Note: The **AUC-ROC** curve (Area Under the Receiver Operating Characteristic Curve) is a performance measurement for classification models. The ROC curve plots the **True Positive Rate** (Sensitivity) against the **False Positive Rate** (1 - Specificity) at various threshold settings. The AUC represents the degree of separability between the classes, with a value of 1 indicating perfect classification and 0.5 representing a model no better than random guessing. A higher AUC indicates better model performance, as it shows that the model distinguishes well between positive and negative classes.

***Fig 5.4*** *AUC-ROC curve*

## Cluster Analysis - Comprehensive Technique for Customer Segmentation

Cluster analysis was employed as the second data analytics technique to group customers into distinct segments based on their demographic, policy, and behavioral attributes. This method is pivotal for personalized marketing and strategic decision-making. By identifying distinct customer clusters, the insurance company can tailor its communication strategies, design customized insurance plans, and improve customer engagement. This segmentation aligns with the business goal of predicting customer interest in vehicle insurance and effectively targeting the right audience.

### Purpose of Cluster Analysis:

For this project, cluster analysis aimed at:

* + - 1. Identify distinct customer groups based on demographic, policy, and behavioral attributes.
      2. Enable targeted marketing campaigns and improve customer engagement.
      3. Provide actionable insights to optimize resource allocation and business strategies.

### Implementation Steps

* + - 1. **Data Preprocessing**:
         * **Feature Selection**: Selected features relevant for segmentation, such as Age, *Vehicle\_Age*, *Annual\_Premium, Vehicle\_Damage*, and *Policy\_Sales\_Channel*.
         * **Feature Scaling**: Applied StandardScaler to standardize continuous features, ensuring that attributes with larger ranges do not dominate the clustering process.
         * **Feature Vector Creation**: Used PySpark's VectorAssembler to combine selected features into a single vector column named features.
      2. **Dimensionality Reduction**:
         * Principal Component Analysis **(**PCA) was employed to reduce the high-dimensional feature space to two principal components for better visualization of clusters. This dimensionality reduction ensured computational efficiency and interpretability while retaining most of the variance in the data.
      3. **K-Means Clustering**:
         * **Algorithm**: K-Means clustering was selected due to its efficiency in handling large datasets and its ability to produce well-defined clusters.
         * **Number of Clusters (k)**: The optimal number of clusters (k=5) was determined using the elbow method and domain knowledge.
         * The K-Means algorithm was applied to the standardized feature vector to group customers into five clusters. PySpark’s KMeans class was used for clustering, and the seed parameter was set for reproducibility.
      4. **Evaluation Metric**:
         * **Silhouette Score**: The silhouette score was used to evaluate clustering performance. It measures how well each data point fits within its assigned cluster compared to other clusters. The model achieved a silhouette score of 0.8277, indicating that the clusters are well-separated and distinct.

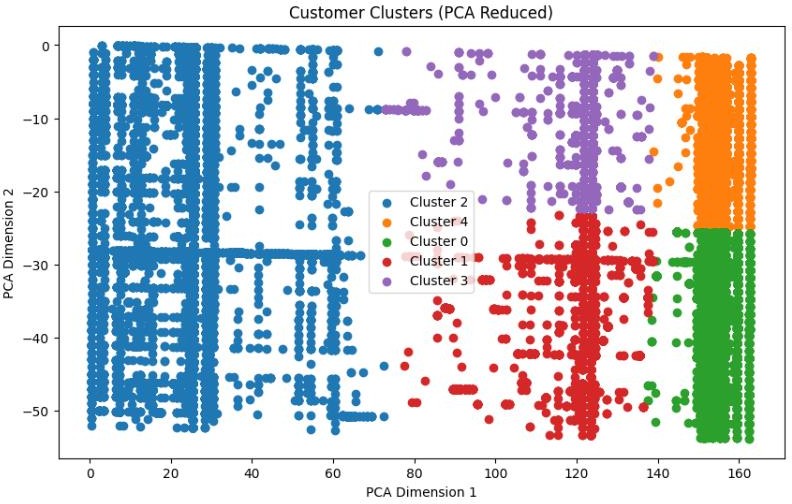
### Importance of Silhouette Score

The Silhouette Score is a crucial metric for evaluating clustering performance, as it measures both cluster cohesion and separation. A high silhouette score signifies tight cohesion within clusters (low intra-cluster distance) and clear separation between clusters (high inter-cluster distance), ensuring that clusters are well- defined and distinct. This score is particularly useful for determining the optimal number of clusters (k) in algorithms like K-Means by calculating the silhouette score for different values of k and selecting the one with the highest score, which often corresponds to the most suitable clustering configuration. Additionally, the silhouette score aids in cluster interpretation by validating that data points within each cluster are more like each other than two points in other clusters, ensuring meaningful and interpretable clusters. Such well-defined clusters are essential for deriving actionable business insights.

The clustering model achieved a silhouette score of **0.8277**, which is close to the ideal score of 1. This high score indicates that the clusters are well-separated, and data points are strongly associated with their respective clusters. The well-defined clusters enable reliable interpretations for targeted marketing and resource allocation.

### PCA Scatter Plot

The clusters were visualized using a scatter plot, with two principal components (PCA dimensions) on the axes. The plot shows clear boundaries between clusters, validating the effectiveness of the K-Means algorithm.



Note: The scatter plot of customer clusters, reduced to two dimensions using Principal Component Analysis (PCA), provides a visual representation of the clusters identified through K-Means clustering. Each cluster corresponds to a distinct group of customers with shared characteristics, enabling targeted segmentation. The clear separation between clusters highlights the algorithm's effectiveness in identifying meaningful groupings. This visualization aids in interpreting customer behaviors and supports the development of personalized marketing strategies.

***Fig 5.5*** *PCA – based Cluster Scatterplot*

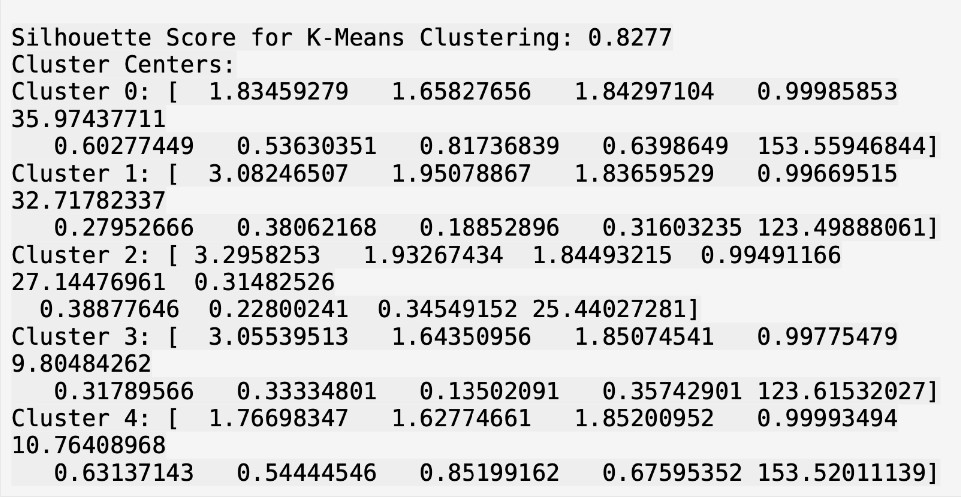
### Cluster Characteristics Extraction

The process of extracting and interpreting cluster characteristics is crucial to understanding the unique attributes of customer segments generated by the clustering algorithm. For this project, the defining traits of each cluster were derived using the cluster centroids, which represent the average values of features within each cluster. These characteristics provide actionable insights for creating targeted strategies.

**Cluster Centers (Centroids)**

Cluster centers are the mean values of all features for the data points belonging to each cluster. They provide a summarized representation of the typical attributes of customers within a cluster. For example, in a dataset containing features like age, annual premiums, vehicle damage history, and sales channel interaction, the centroid of a cluster reflects the average profile of the customers in that cluster.

After applying the K-Means clustering algorithm, the centroids were calculated as part of the output. These values were extracted for further analysis to understand the differences and similarities between clusters. The cluster centers were expressed numerically (e.g., average age, scaled values for annual premium), making it easier to interpret the general behavior or traits of each group.



***Fig 5.6*** *Cluster Silhouette Score Output*

The defining characteristics of each cluster were derived by analyzing the centroid values.

* + - 1. **Age Distribution**:
         * By examining the average age values, the clusters were categorized into segments such as younger, middle-aged, or older customers.
         * For example:

Cluster 0 may represent older customers (e.g., average age ~60).

Cluster 3 may represent younger customers (e.g., average age ~25).

* + - 1. **Annual Premium**:
         * The centroids provided insights into the spending patterns of customers.
         * For instance:

Clusters with high average annual premiums were interpreted as high-value customers willing to invest in comprehensive policies.

Clusters with lower premiums indicated cost-sensitive customers prioritizing affordability.

* + - 1. **Vehicle Damage History**:
         * The proportion of customers with a history of vehicle damage (e.g., values close to 1) was analyzed to identify risk-prone clusters.
         * Clusters with a higher proportion of vehicle damage history indicated customers who may perceive higher value in vehicle insurance due to their past experiences.
      2. **Previously Insured**:
         * The cluster centers showed the proportion of customers who already had vehicle insurance.
         * Clusters with a higher percentage of "Previously Insured" (close to 1) suggested existing policyholders, likely interested in policy upgrades or renewals.
         * Clusters, with a lower percentage indicated potential new customers or first-time buyers.
      3. **Policy Sales Channel**:
         * The interaction levels with different sales channels were analyzed.
         * For example, some clusters showed higher reliance on digital channels (higher values for specific sales codes), while others interacted more through in-person agents.
      4. **Vintage (Customer Tenure)**:
         * The average "vintage" values indicated how long customers had been associated with the company.
         * Clusters with high vintage represented loyal, long-term customers.
         * Low vintage clusters indicated newer customers, potentially requiring different engagement strategies.

**Identified Clusters**

The analysis identified five unique clusters, differentiated by features such as age, annual premiums, vehicle damage history, and sales channel interaction. The clusters were identified as follows:

* + - * + **Cluster 0 (Cost-Conscious Customers)**: Customers with low premiums and older vehicles, focusing on affordability.
        + **Cluster 1 (Premium Buyers)**: Younger customers with newer vehicles and higher premiums, interested in comprehensive coverage.
        + **Cluster 2 (Risk-Averse Customers)**: Previously insured individuals, cautious and focused on maintaining coverage.
        + **Cluster 3 (First-Time Buyers)**: New customers with limited prior insurance experience, needing guidance and education.
        + **Cluster 4 (Loyalists)**: Long-standing customers with moderate premiums and high vintage, presenting opportunities for upselling.

# Model Interpretation

The interpretation of the models provided actionable insights into customer behavior and key factors influencing the likelihood of purchasing additional insurance. These insights supported the development of targeted marketing strategies and enhanced decision-making for business growth.

* 1. **Insights from Feature Importance**

The XGBoost model's feature importance analysis highlighted the most influential predictors, offering valuable business insights:

* + 1. **Vehicle\_Damage\_Indexed**:
       - **Business Insight**: Customers with a history of vehicle damage are significantly more likely to purchase additional insurance products. This aligns with the understanding that customers who have experienced past losses are more risk-aware and willing to invest in protection.
       - **Actionable Strategy**: Focus marketing efforts on customers with prior claims or vehicle damage history. Offer tailored insurance packages emphasizing enhanced protection against potential future losses.
    2. **Vehicle\_Age\_Indexed**:
       - **Business Insight**: Newer vehicles were strongly associated with higher insurance adoption rates, as customers with newer vehicles often perceive more value in safeguarding their high-value assets.
       - **Actionable Strategy**: Target owners of recently purchased or newer vehicles with comprehensive insurance plans and value-added services, such as maintenance coverage or theft protection.
    3. **Gender\_Indexed**:
       - **Business Insight**: Gender differences in purchasing behavior were observed, indicating distinct preferences and insurance needs based on demographic trends.
       - **Actionable Strategy**: Use demographic segmentation to design gender-specific campaigns, ensuring personalized communication to address unique concerns and priorities.
    4. **Previously\_Insured**:
       - **Business Insight**: Customers who are already insured with other providers exhibited a lower likelihood of purchasing additional insurance. This suggests a need to differentiate offerings or highlight competitive advantages.
       - **Actionable Strategy**: Develop campaigns focused on the competitive edge of the company’s insurance products (e.g., cost savings, superior coverage) to attract customers insured by competitors.

These features provided a clear understanding of the key drivers influencing customer decisions, enabling a data-driven approach to marketing and sales strategies.

* 1. **Insights from Clustering**

The **K-Means clustering analysis** identified five distinct customer segments, each with unique behavioral and demographic characteristics. These insights supported tailored engagement strategies:

* + 1. **Cluster 0: Cost-Conscious Customers**:
       - **Attributes**: Low premiums, older vehicles, high interaction with policy sales channels.
       - **Business Insight**: This segment is highly price-sensitive, prioritizing affordability over comprehensive coverage.
       - **Actionable Strategy**: Offer budget-friendly insurance plans or promotional discounts to capture this group’s attention.
    2. **Cluster 1: Premium Buyers**:
       - **Attributes**: Younger customers with newer vehicles and higher premiums.
       - **Business Insight**: This group values premium insurance offerings and is willing to invest in better coverage.
       - **Actionable Strategy**: Highlight premium products with extended benefits and convenience features, such as fast claims processing or exclusive add-ons.
    3. **Cluster 2: Risk-Averse Customers**:
       - **Attributes**: Previously insured, consistent vehicle damage claims, cautious behavior.
       - **Business Insight**: These customers prioritize maintaining coverage to mitigate potential risks.
       - **Actionable Strategy**: Retain this segment by emphasizing loyalty programs, long-term savings, and enhanced peace-of-mind messaging.
    4. **Cluster 3: First-Time Buyers**:
       - **Attributes**: Younger individuals with lower interaction with policy channels, recent policy purchases.
       - **Business Insight**: This group is new to insurance and may require guidance to understand the benefits of comprehensive coverage.
       - **Actionable Strategy**: Launch beginner-friendly plans with simple, easy-to-understand features. Provide educational resources to build trust and encourage adoption.
    5. **Cluster 4: Loyalists**:
       - **Attributes**: Long-standing customers with moderate premiums and high vintage.
       - **Business Insight**: Loyal customers represent a stable base for cross-selling and upselling opportunities.
       - **Actionable Strategy**: Strengthen engagement through personalized communication and incentives. Upsell additional policies such as health riders or home insurance.

By identifying these clusters, the business can implement targeted marketing strategies tailored to the specific needs and behaviors of each customer group, improving conversion rates and customer satisfaction.

* 1. **Business Insights from Model Performance**

The performance metrics further validated the applicability of the models in supporting the business:

* + 1. **Precision and Recall**:
       - **Insight**: XGBoost achieved a balance between precision and recall, ensuring that most potential buyers were identified without overwhelming the marketing team with false positives.
       - **Business Benefit**: Reduced wastage of marketing resources while capturing a high proportion of potential customers.
    2. **AUC (Area Under the Curve)**:
       - **Insight**: XGBoost achieved an AUC of **83.63%**, indicating its ability to distinguish between customers likely to purchase insurance and those unlikely to do so.
       - **Business Benefit**: Enabled focused targeting of customers with high purchasing propensity, improving marketing efficiency.
    3. **Model Comparison**:
       - **Insight**: XGBoost outperformed Logistic Regression and Decision Tree Classifier across all key metrics, proving its superiority in handling complex patterns and imbalanced data.
       - **Business Benefit**: Established confidence in the predictive capability of XGBoost, ensuring reliable decision-making.

# Business Implications and Recommendations

The combined insights from predictive modeling and cluster analysis offer a robust foundation for data- driven decision-making in the insurance sector, with actionable strategies to enhance customer engagement, optimize resource allocation, and maximize profitability. These findings emphasize the importance of leveraging advanced analytics to improve marketing efficiency, customer targeting, and operational effectiveness.

* 1. **Business Implications**
     1. **Enhanced Targeting Through Predictive Modeling**:
        + The XGBoost model identified key predictors of insurance purchase behavior, such as *Vehicle\_Damage\_Indexed* and *Vehicle\_Age\_Indexed*. These insights enable the business to target customers more likely to purchase additional insurance with personalized offers and tailored communication strategies. The model’s high precision and recall balance ensures that the marketing team can focus on high-propensity customers while minimizing wasted resources on unlikely buyers.
     2. **Segmentation-Based Marketing with Clustering**:
        + K-Means clustering revealed five distinct customer segments, including "Cost-Conscious Customers," "Premium Buyers," and "First-Time Buyers." Each segment exhibits unique behaviors and preferences, allowing for the development of personalized marketing campaigns.
        + For example, "Premium Buyers" can be targeted with high-value, comprehensive insurance plans, while "First-Time Buyers" can be engaged with beginner-friendly educational materials and simple plans.
     3. **Improved Customer Retention and Loyalty**:
        + Insights from *Previously\_Insured* and the "Loyalist" cluster highlight opportunities to retain long-standing customers by offering loyalty rewards, personalized renewal plans, or value- added services such as extended coverage and discounts.1
     4. **Informed Product Development**:
        + By understanding the key features driving customer behavior (e.g., *Vehicle\_Age\_Indexed*), the business can prioritize the development of products that cater to customer needs, such as customized policies for owners of newer vehicles or packages addressing specific risk factors.2
     5. **Operational Efficiency**:
        + The integration of predictive modeling and clustering into the marketing pipeline streamlines operations by enabling data-driven decision-making. Teams can allocate resources more efficiently, focusing on customers and segments with the highest revenue potential.3

1.Meyer-Waarden, L. (2007). *The impact of loyalty programs on repeat purchase behavior*. Journal of Marketing Research, 44(2), 67-78. Retrieved from [https://hbr.org](https://hbr.org/)

2 McKinsey & Company. (2023). *Next-level insurance product development*. Retrieved from [https://www.mckinsey.com](https://www.mckinsey.com/)

3. Salesforce. (2023). *Integrating CRM systems for efficient customer management*. Retrieved from [https://www.salesforce.com](https://www.salesforce.com/resources/articles/integrate-crm/)

* 1. **Recommendations**
     1. **Targeted Marketing Strategies**:
        + Leverage the predictive insights from XGBoost to create targeted campaigns for high-priority customers, such as those with past vehicle damage or newer vehicles.
        + Use demographic segmentation (e.g., gender, age) to design tailored messaging and offers that resonate with specific customer groups.
     2. **Segment-Specific Engagement Plans**:
        + For "Cost-Conscious Customers," emphasize affordability by offering bundled discounts or promotional rates.
        + Engage "Premium Buyers" with premium-tier products and exclusive benefits, such as expedited claims processing or enhanced coverage options.
        + For "First-Time Buyers," focuses on educational outreach and beginner-friendly plans to build trust and encourage adoption.
     3. **Loyalty and Retention Programs**:
        + Introduce reward programs for "Loyalists" and "Risk-Averse Customers" to maintain their long-term engagement. For example, offer discounts on renewals or additional policies as a token of appreciation for their loyalty.
        + Implement feedback mechanisms to understand their evolving needs and refine offerings accordingly.
     4. **Expand Cross-Selling Opportunities**:
        + Use the clustering insights to identify cross-selling opportunities. For example, offer health or home insurance packages to customers in the "Loyalist" or "Risk-Averse" segments.
        + Promote complementary services or add-ons, such as roadside assistance or accident coverage, tailored to each cluster’s preferences.
     5. **Continuous Model Improvement**:
        + Retrain and update the predictive models and clustering algorithms with new data to adapt to changing customer behaviors and market trends regularly.
        + Monitor the performance of marketing campaigns using A/B testing to validate the effectiveness of the recommendations derived from the models.
     6. **Visual Dashboards for Decision-Making**:
        + Develop interactive dashboards to visualize key metrics, feature importance, and cluster characteristics. This will empower marketing and sales teams to make data-driven decisions in real time.
     7. **Adopt Technology for Scalability**:
        + Integrate the predictive and clustering models into an automated customer relationship management (CRM) system. This will enable seamless execution of marketing campaigns and provide real-time insights into customer engagement and conversion.
  2. **Expected Business Outcomes**
     1. **Increased Revenue**:
        + By focusing on high-propensity customers and tailoring marketing strategies, the business can expect higher conversion rates and improved cross-selling success.
     2. **Improved Customer Satisfaction**:
        + Personalized engagement strategies ensure that customers feel understood and valued, leading to greater satisfaction and loyalty.
     3. **Cost Optimization**:
        + Efficient resource allocation reduces the cost of marketing while improving ROI by targeting customers more likely to purchase.
     4. **Scalable and Data-Driven Operations**:
        + Integrating predictive analytics and clustering into business workflows enables scalable and adaptive strategies, ensuring long-term competitiveness in the market.

This analysis demonstrates how the insights derived from the models and clusters can be translated into tangible business benefits. By implementing these recommendations, the company can enhance its marketing effectiveness, deepen customer relationships, and achieve sustainable growth.

# 8. Future Scope of Improvement

While the project achieved significant results, there are opportunities to enhance its effectiveness by incorporating additional data, refining methodologies, and leveraging advanced technologies. Below are the key areas for improvement, with a focus on prioritized improvements:

1. **Enhanced Features**

Adding more customer-centric data points can significantly improve the predictive power and segmentation accuracy:

* + Income and Credit Score: Including these variables would provide insights into financial capacity and risk, leading to more precise targeting.
  + Driving History: Historical driving patterns and accident data could help refine risk predictions.
  + Real-Time Data: Integrating real-time customer interactions, such as website visits or call logs, could enhance dynamic predictions.

1. **Feature Selection**

Refining feature inputs can eliminate noise and improve model efficiency:

* + Use statistical techniques like correlation analysis, PCA (Principal Component Analysis), and mutual information to identify the most relevant features.
  + Conduct domain-specific analysis to ensure selected features align with business goals.

1. **Hyperparameter Tuning**

Optimizing model parameters is critical for improving accuracy and reducing bias:

* + Implement advanced tuning techniques, such as Bayesian optimization or random search, to automate the parameter selection process.
  + Evaluate performance across different metrics (e.g., precision, recall, F1-score) to ensure balanced optimization.

1. **Dynamic Updates**

To ensure the model adapts to changing customer behavior:

* + Regularly retrain the model with updated datasets, incorporating seasonal trends and market shifts.
  + Use incremental learning methods for real-time data updates without full retraining.

1. **Improved Interpretability**

Building trust among stakeholders is vital, particularly in regulated industries like insurance:

* + Leverage tools like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model- Agnostic Explanations) to provide clear explanations for model predictions.
  + Develop intuitive dashboards to communicate insights effectively.

Other possibilities for improvement include integrating advanced algorithms such as neural networks or transformers to capture complex, non-linear patterns in customer data. Enhanced clustering techniques, like DBSCAN or hierarchical clustering, could identify overlapping or irregular data segments, offering deeper insights. Incorporating temporal analysis, such as time-series models, would help understand seasonal trends and long-term customer behavior. Real-time deployment through streaming data pipelines, like Apache Kafka, could ensure dynamic updates to predictions and clustering results. Additionally, A/B testing could be used to validate marketing strategies and refine predictive insights, while cross-selling and upselling opportunities could be better identified by leveraging segmentation insights to design bundled products or add-on services, such as home or health insurance policies. These enhancements could significantly elevate the project’s scalability, adaptability, and business impact.

# 9. Conclusion

This project successfully leveraged advanced data analytics techniques, including predictive modeling and clustering, to address the business problem of predicting customer interest in vehicle insurance. Through a combination of XGBoost, Logistic Regression, and Decision Tree models, we identified key factors influencing customer decisions, such as vehicle damage, vehicle age, and insurance history. XGBoost emerged as the most effective model, with high accuracy and feature importance insights that guided targeted marketing strategies.

Cluster analysis further enriched our understanding by segmenting customers into distinct groups, each with unique characteristics. This allowed for the development of tailored marketing campaigns, optimizing resource allocation and enhancing customer engagement. Key findings highlighted the potential to focus on high-interest customers, re-engage low-interest ones through education, and cross- sell policies to maximize revenue.

The importance of this analysis lies in its ability to make the insurance company's marketing strategies more data-driven and effective. By identifying specific customer traits and preferences, the company can allocate resources efficiently, increase conversions, and improve overall profitability. The integration of these insights into decision-making processes ensures a competitive edge in a dynamic market.

# 10 Additional References

Colab Notebook Link – [Click here!](https://colab.research.google.com/drive/1tY1FM47ql-YD78_5Wna_mZ2p4g7xca9S?usp=sharing)

1. Kotler, P., & Keller, K. L. (2016). *Marketing Management* (15th ed.). Pearson Education.
   * **In-text citation**: (Kotler & Keller, 2016)
   * **Relevance**: This book provides foundational principles on targeted marketing strategies, essential for designing campaigns based on customer segmentation.
2. Homburg, C., Schäfer, H., & Schneider, J. (2012). *Sales Excellence: Systematic Sales Management*. Springer.
   * **In-text citation**: (Homburg, Schäfer, & Schneider, 2012)
   * **Relevance**: Explains how predictive modeling can enhance sales efforts and resource allocation, directly applicable to targeting high-probability insurance buyers.
3. Insure.com. (2023). *Understanding the Role of Vehicle Damage in Insurance Premiums*. Retrieved from [Insure.com](https://www.insure.com/)
   * **In-text citation**: (Insure.com, 2023)
   * **Relevance**: Highlights the impact of vehicle damage history on insurance pricing and customer behavior, aligning with feature importance findings in this project.
4. Bhat, S., & Darzi, M. A. (2016). *Customer relationship management: A key to sustainable growth in the insurance industry*. *Pacific Business Review International, 9*(1), 54–63.
   * **In-text citation**: (Bhat & Darzi, 2016)
   * **Relevance**: Discusses CRM strategies and how data analytics can improve customer targeting and retention, relevant to the business implications derived from the project.
5. Statista. (2023). *Global Insurance Market Size and Growth Trends*. Retrieved from [Statista](https://www.statista.com/)
   * **In-text citation**: (Statista, 2023)
   * **Relevance**: Provides industry insights into growth opportunities, emphasizing the importance of cross-selling health and vehicle insurance policies.
6. Association for Insurance Marketing and Analytics (AIMA). (2022). *Innovative Marketing Strategies in the Insurance Sector*. Retrieved from [AIMA](https://www.aima.org/)
   * **In-text citation**: (AIMA, 2022)
   * **Relevance**: Highlights how marketing analytics and segmentation can improve profitability in the insurance industry.

ChatGPT was utilized to structure, refine, and enhance the project report and presentation. It provided guidance on summarizing data analytics techniques, crafting professional narratives, and ensuring clear explanations of complex topics such as predictive modeling and clustering. Additionally, it assisted in generating insights and actionable business implications based on the project's findings.