# **MSC 641** Customer Segmentation Final Report

By: Sahith Akumadupula, Jordyn Gaither, Manju Bhargavi Kurapati, Durga Koushik Perla

**Table of Contents**

1. Introduction
2. Data Description and Exploration
3. Methods and Results

* Feature Engineering
* Decision Tree
* Random Forest Classification
* Gradient Boosting
* XGBoost Classifier
* Support Vector Machine (SVM)

1. Conclusion

* Recommendations

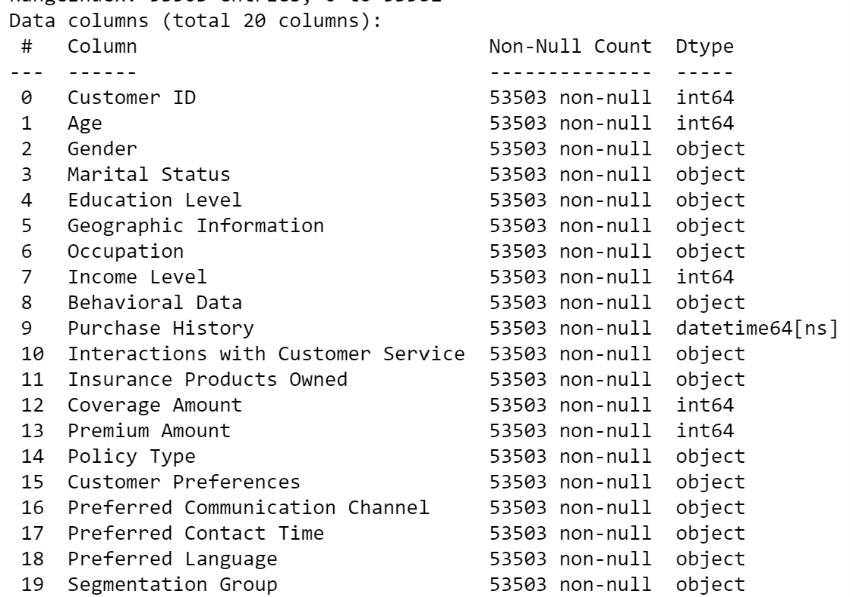
## Introduction

In the context of the insurance industry, understanding customer behavior and preferences is crucial for tailoring products and services effectively. This project aims to analyze customer data from an insurance company, segment customers based on behavior and characteristics, and develop predictive models to forecast insurance coverage amounts and policy types. By doing so, we aim to provide actionable insights for the insurance company to enhance customer targeting and improve forecasting capabilities.

## Data Description and Exploration

**Dataset Overview**

The dataset comprises 53,503 rows and 20 columns, consisting of date-time, integer, and categorical (object) variables. There are no missing, null, or duplicate values in the dataset. Each variable represents specific aspects of customer behavior, demographics, and interactions with insurance products.

****A screenshot of a computer

Description automatically generated

**Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) was conducted to gain insights into the dataset:

* **Data Types and Structure:** Identified the types of variables and their distribution.
* **Unique Values per Variable:** Investigated the range and diversity of values in each column.
* **Distribution Analysis:** Examined the distribution of numeric variables and explored categorical variable distributions.
* **Relationship Analysis:** Investigated correlations between variables and potential patterns

A collage of graphs

Description automatically generated

## Methods and Results

* **Feature Engineering**
* **Decision Tree**

A Decision Tree classifier was trained and evaluated, but it exhibited limited discriminatory power with low accuracy and AUC score. This suggests that decision tree-based models may not be suitable for this dataset without additional feature engineering or model optimization.

**Key Metrics**: Accuracy: 0.2681992337164751  
Confusion Matrix:  
[[ 748  628  935  494]  
 [ 636  560  812  427]  
 [ 932  867 1247  652]  
 [ 443  427  578  315]]  
AUC Score: 0.5021665947572699

A screenshot of a computer

Description automatically generated

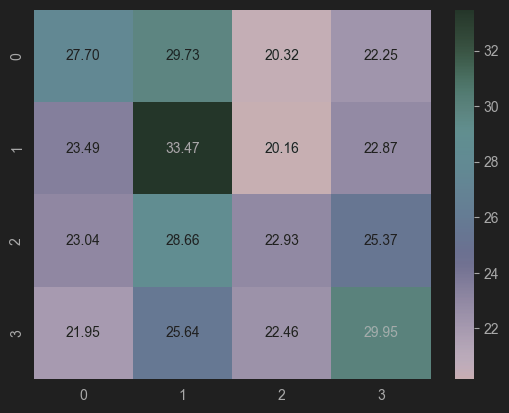
**A graph showing the amount of a number of companies

Description automatically generated with medium confidence**

* **Random Forest Classification**

We employed Random Forest, an ensemble method, to improve predictive performance. By optimizing hyperparameters using GridSearchCV, we achieved better accuracy and performance compared to individual decision trees.

**Key Metrics:**Best hyperparameters: {'max\_depth': 8, 'n\_estimators': 30}  
Test accuracy: 0.277357256331184  
Train accuracy: 0.46568661125231847  
Confusion Matrix:  
[[ 777  834  570  624]  
 [ 572  815  491  557]  
 [ 852 1060  848  938]  
 [ 387  452  396  528]]  
False Negative Rate (FNR): 0.4124008651766402  
False Positive Rate (FPR): 0.5176908752327747  
Misclassification Rate (MR): 0.46897931954636424



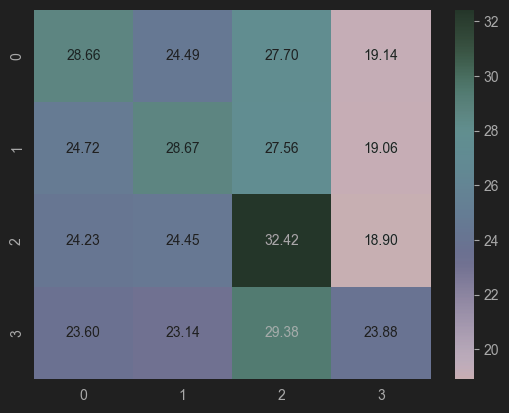
* **Gradient Boosting**

Despite achieving a modest accuracy of approximately 35.32%, Gradient Boosting showed limitations in accurately predicting certain segmentation groups, particularly Group 4. Further analysis and feature engineering are recommended to enhance model performance.

**XGBoost Classifier**

The XGBoost Classifier demonstrated moderate predictive ability with an accuracy of 29.17% and an AUC score of 0.543. While the classifier shows promise, additional refinement and parameter tuning are necessary to optimize performance.

Best model accuracy on test data: 29.52%  
Best parameters found:  {'learning\_rate': 0.3, 'max\_depth': 7}  
XGBoost Classifier Confusion Matrix:  
[[ 807  681  887  430]  
 [ 606  641  768  420]  
 [ 926  846 1352  574]  
 [ 435  386  583  359]]  
XGBoost Classifier AUC Score: 0.5311887302005017

****

* **Support Vector Machine (SVM)**

SVM was employed to predict insurance policy types based on customer attributes. We experimented with different kernels including linear, polynomial (degrees 2, 3, 4), and radial basis function (RBF). The poly kernel demonstrated the best performance based on our evaluation metrics.

**Evaluation Metrics:**

The model was evaluated on the test set using accuracy and confusion matrix.

Insights:

A screenshot of a computer

Description automatically generated

Confusion Matrix:

[[  10  262   33  148 1331]

 [  14  456   41  176 1631]

 [  10  328   42  135 1332]

 [  14  397   27  202 1337]

 [  10  394   41  170 2160]]

Test accuracy for SVM: 0.2681992337164751

False Negative Rate (FNR): 0.029787234042553193

False Positive Rate (FPR): 0.9632352941176471

Misclassification Rate (MR): 0.3719676549865229

The poly kernel was selected as the best-performing kernel for this dataset.

However, the model's performance, as indicated by accuracy and confusion matrix, was unsatisfactory, suggesting the need for further optimization or alternative models.

## Conclusion

In conclusion, our analysis highlights the challenges in accurately predicting customer segmentation using machine learning models on the provided dataset. Despite efforts with various algorithms, predictive accuracy remains modest. For practical applications such as marketing analysis, businesses may benefit from tailored segmentation strategies based on domain knowledge and specific customer insights.

**Recommendations:** Based on our findings, we recommend the following:

* **Feature Engineering:** Explore additional features or transformations to enhance model performance.
* **Model Optimization:** Conduct further parameter tuning and experimentation with advanced algorithms.
* **Domain-Specific Segmentation:** Utilize domain knowledge to create customized segmentation strategies aligned with business objectives.

By implementing these recommendations, businesses can leverage customer segmentation effectively to improve marketing strategies and enhance customer satisfaction.