Market Segmentation Project

• • •

By,
Jesse R. Mulwa

Project Overview

Objective: Classify customers into distinct segments based on various features using machine learning algorithms.

Algorithms Used:

- Support Vector Machine (SVM)
- Logistic Regression
- Decision Tree

Data: Customer data with numerical and categorical features, targeting segments like APAC, Africa, US, EU, etc.

Understanding the problem

Clear Business Goals

Align segmentation with clear business goals for actionable insights.

Customer Value

Prioritize high-value customer segments to optimize strategies.

Data Insights

Translate technical results into meaningful, business-relevant insights.

DataSet Overview:

Features: Customer attributes like spending, region, demographic data, etc.

Target Variable: Customer segments (e.g., APAC, Africa, EU, LATAM)

Preprocessing:

- Categorical feature encoding using one-hot encoding.
- Scaling numerical features for better performance with SVM.
- Handling imbalanced data using `class_weight='balanced'`.

Understanding Model Selection

Logistic Regression:

- Simple and interpretable.
- Handles linear decision boundaries well.
- Works well with balanced data and `class_weight='balanced'` for imbalance.

Decision Tree:

- Captures non-linear relationships.
- Prone to overfitting without pruning.
- Requires careful hyperparameter tuning.

Support Vector Machine (SVM):

- Best for high-dimensional data.
- Handles linear and non-linear separability using kernel tricks.
- `class_weight='balanced'` used to address class imbalance.

Model Evaluation Metrics

Metrics:

- **Precision**: Ability to identify positive instances.
- **Recall**: Ability to detect all relevant instances.
- **F1-Score**: Harmonic mean of precision and recall.
- **Accuracy**: Overall correct predictions.

Evaluation Approach:

- Classification report to compare precision, recall, and F1-score.
- Confusion matrix to visualize misclassifications.

Market trends

Trend 1

Developed markets have disproportionately more products marked as high priority than developing markets.

Client Implications:
- Service delivery is less reliable in developing markets.
- The data shows prevalence of unethical segmentation across ethnic, racial or regional lines.

Trend 2

Improved data collection and management across all markets.

Client Implications:
- Better data-driven protocols and processes leading to customer satisfaction
- Opening up of markets that were previously inaccessible or unjustifiable for commerce.

Process analysis

Challenges & Findings

- **Imbalanced Data**: Affects the performance of all models, especially for minority classes.

- **Model Overfitting**: Decision Trees prone to overfitting without proper pruning.

- **Feature Selection**: Identifying the most relevant features for improved performance

20
5
15

29
4
25

39
4
35

27
5
22

20XX   20XX   20XX   20XX
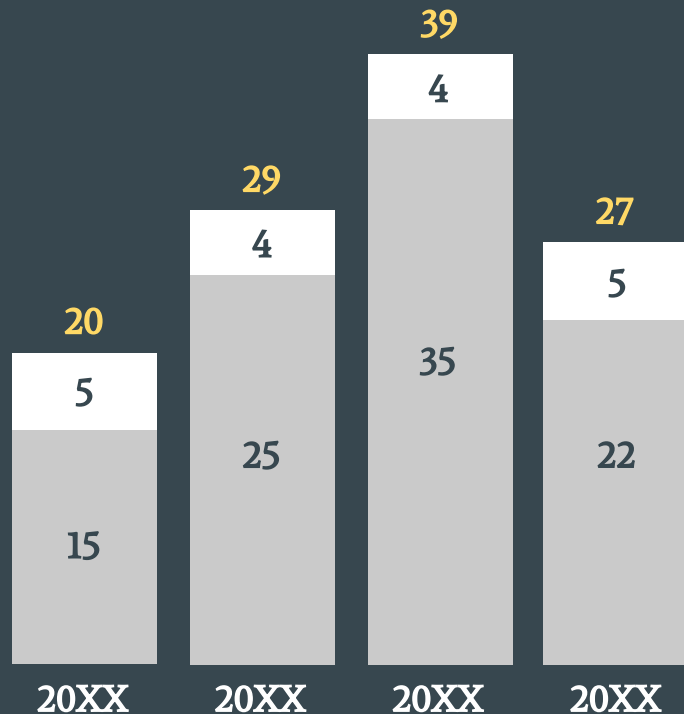
Main Takeaways

SVM:
Performed well enough but struggled with minority classes.

Logistic Regression:
Is a strong baseline with easy interpretation.

Decision Tree:
Requires careful tuning to avoid overfitting and improve generalization.

Imbalance Handling:
Class weight adjustment and oversampling are essential for improving model performance on imbalanced datasets.

Timeline & Next Steps

Improvement:
Explore ensemble methods like Random Forest and XGBoost for better performance.

Optimization:
Hyperparameter tuning for SVM, Decision Tree, and Logistic Regression.

Deployment:
Develop an automated pipeline for real-time customer segmentation.

The End.
Thank you.