Complete Capstone Project

Predicting customer behavior using machine learning algorithms: Customer Insight segmentation App

Abadit Weldeslassie

DSC-590: Data Science Capstone Project

DSC-560: Dr. Brian Letort

October 13, 2024

Table of Contents

- Overview of design concepts
- Data acquisition
- Data preparation
- Data exploratory analysis
- Model selection
- Model training and evaluation
- Result interpretation and business insights
- Detailed model pipeline design
- Stages of model pipeline
- Data sources
- Data types and formatting
- Data model
- Methodology
- Infrastructure and environment configuration
- Security
- References

3. **Milestone 3: Implementation**
    - System entities
    - Functional requirements
    - Implementation plan
    - Introduction
    - System requirements
    - Data overview
    - Step by step instructions for using the product
    - Conclusion

4. **Milestone 4: Results Analysis or Testing Components**
    - Component testing
    - Requirement testing
    - Functional requirements
    - Nonfunctional requirements
    - System testing
    - User guide
    - System administration guide
5. **Conclusion**

1. Milestone 1: Project Proposal and Requirement Analysis

Project overview

1. Introduction

Understanding what customers want and anticipating their needs is interesting and a constant challenge for business. Machine learning and predictive analysis are the most required tools for business to uncover and understand customer needs from huge collection of data. Predictive analytics involves certain manipulations on the existing dataset with the goal of identifying new trends and patterns to predict future outcomes and trends. With help of machine learning algorithms, researchers address some problems in customer behavior analysis. Machine learning generates predictions for individual customers and those predictions can drive how each customer is served. As a result, decision makers increasingly recognize that ML can have a huge impact on the customer experience and begin to focus generating concrete value with machine learning and accelerating and expanding its use.

2. Problem statement

The objective of this study is to provide in-depth predictive analysis of customer behavior using machine learning algorithms which plays a vital role in decision making, improve profit rates of business, increase customer satisfaction, and reduce risk by identifying them at the early stage.

3. Research question/Hypotheses:

H1: how is customer behavior highly related to customer experience within the business?

H2: Is there any significant relationship between customer churn, customer experience and the success of the business?

4. Background and context

According to the literature review of google scholars, ScienceDirect, and IEEE Xplore I was able to explore and understand how and why customers often shop differently and may have different needs or different values to the business. Most of previous studies focus on specific customer behavior such as customer purchase behavior analysis, customer churn analysis using several machine learning algorithms such as, decision tree, random forest, and support vector to understand trends and patterns based on historical data. However, those models often fall short in comprehensive analysis of customer behavior including their demography, geography and behavioral analysis. Building upon those previous studies this study aims to delve deeper into the identification of vital yet unnoticed variables which are essential to increase the success rate of acquiring customers, sales and establishes a sense of competitiveness in market.

The project benefits business, decision makers and customers as accurate predictive analysis such as customer segmentation can lead to an effective marketing strategy which in turn can result in better customer satisfaction and greater revenue. Knowledge of how to predict customer behavior is essential in today's market both for enhancing customer retention and business growth. Some of the benefits of Predictive analysis and customer segmentation are

precise segmentation of audiences, thus helps business to define the actual value of a customer and then make an efficient decision on whether it is worth investing effort and time to retain the customer. Personalized marketing experience, if a business understands its customers and factors impacting their behaviors, they can create customer experience that satisfied them which led to customer loyalty and retention.

Project objectives

Customer behavior analysis is useful for an enterprise, it helps business to better understand their customer needs, preferences and buying pattern, thus helps marketing team to tailor their efforts to reach out customers in the most efficient way for a better customer satisfaction and marketing strategies.

The purpose of this project is to predict customer behavior using machine learning algorithms to make data driven decisions regarding marketing strategies and customer satisfaction. Nowadays the use of machine learning methods to predict customer behavior analysis has become more attractive and some of the benefits and opportunities of predicting customer behavior using machine learning algorithms are:

- get an insightful information about the customers helps to create a personalized connection with customers to sell and market products, services effectively
- improve customer engagement: using customer behavior data and purchasing history, business can create tailored and targeted customer engagement strategies

- optimized marketing spends targeting specific segments helps to allocate budget more efficiently, avoiding wasteful spending and thus yields higher return on investment
- informed choices: understanding customers provides data-driven insights that guide strategic decisions, minimizing guesswork thus leading to more effective decision making and better outcomes.

Project scope

This project aims to develop a machine learning model to predict customer behavior, with a focus on improving customer retention and enhancing targeted marketing strategies. The model will be built using customer demographic and behavioral information. The project is designed to provide actionable insights that help businesses better understand their customers and make data-driven decisions.

Work breakdown structure

The following tasks are required to satisfy the project objectives:

1. **Literature review**

Utilizing the existing literature, I explore and understand the factors that affect the experience and satisfaction of customers within the business. The most useful database for my search was google scholar, ScienceDirect, and IEEE Xplore. According to the search result I was able to explore and understand how and why customers often shop differently and may have different needs or different values to the business.

2. **Data collection and preprocessing**

   - Collect data from relevant sources (Kaggle, UCI)

- Clean and preprocess data (handling missing values, normalization, feature engineering)
- Split data into training and testing set

3. **Model Development**
   - Select appropriate machine learning algorithms (e.g., decision trees, random forests, logistic regression, K-NN, support vector machine)
   - Train and validate models using the training and testing sets
   - Fine-tune model parameters to optimize performance

4. **Model Evaluation**
   - Evaluate model performance using metrics such as accuracy, precision, recall, and F1-score
   - Compare models and select the best-performing one for deployment
   - documentation and reporting: documenting the process, findings, and analysis of the research.

5. **User Interface Development**
   - Design and develop a dashboard or interface for stakeholders to interact with the model
   - Integrate data visualization tools to display insights and predictions
   - Implement a feature for exporting reports and data

6. **Stakeholders:**

- Data science team: data scientist, responsible for data collection, preprocessing, model development, and evaluation
- End users: sales and Marketing Executives, Utilize the insights generated from the model for decision-making
- Advisors: instructors, provides project oversight, ensures alignment with business goals
- peer providing guidance and feedback

7. **resources required**

- data resources: access to customers data, demographic and behavioral information
- machine learning and analytical software (python, pandas, NumPy, matplotlib, seaborn, scikit learn)
- cloud computing resources (cloud computing services for model training and deployment)

Project completion

To determine the success of the project several key measures will be used. These measures ensure that the project objectives are met and that the deliverables align with stakeholder expectations.

- **Project success measures**

1. **Model performance metrics:** evaluate the model using the following performance metrics:

- **Accuracy:** one of the most straightforward metrics used to evaluate the performance of a classification model. It measures the proportion of correctly classified instances (both true positives and true negatives) out of the total number of instances.
- **Classification report:** detailed summary of the performance of a classification model, providing insights into how well the model is performing across various metrics for each class
- **confusion matrix:** used to evaluate the performance of a classification model by showing the actual vs. predicted classifications. It provides a more detailed breakdown of model performance than accuracy alone, making it especially useful for understanding how well your model is distinguishing between different classes.

2. **Business impact:** the model will help businesses in three aspects:
    - personalized marketing: the product will help the sales team to create highly targeted marketing campaigns focused on the interest and behaviors of specific customer groups, leading to a more satisfying customer experience
    - Efficient resource allocation: business can focus their marketing, sales, and product development effort on segments that respond positively which maximize revenue Increase
    - customer retention: by meeting the specific needs of each segment business can build relationships with customers, increasing loyalty and reducing churn

- **completion criteria**
    - data quality and preprocessing: ensuring that the data is complete and consistency by handling missing value, outliers and any inconsistencies in the data

  - model validation and testing: use reliable model validation techniques to ensure the model's generalizability to new data

- **Assumptions and constraints**
  - Data availability: the assumption that sufficient historical data will be available.
  - Model generalizability assumes that the model trained on historical data will perform well on future data.
  - Data constraints: limited access to data due to privacy concerns or data collection limitations.
  - Time constraints: deadlines that may limit the scope of model tuning and optimization.

- **Explicit Goals and Their Relative Importance**
  - Develop an accurate predictive model: the core objective is to develop an accurate model that can predict customer segments with at least 60% accuracy.
  - Improve business decision: ensuring that the model's prediction led to actionable insights which enhance business outcomes.

Project controls

1. **Risk management**

- **Technical risks:**
  - **Data quality issues:** data might me incomplete, inaccurate, or biased which can affect model performance
  - **Preventative Steps:** Implement robust data cleaning procedures, and regularly update datasets

  - **Model Performance:** The chosen algorithms may not generalize well to unseen data or might overfit.
  - **Preventative Steps:** Use cross-validation, hyperparameter tuning, and explore different algorithms to find the best fit.

- **Operational risks**
  - **Resource Constraints:** limited computational resources might slow down the development process.
  - **Preventative Steps:** optimize resource usage, consider cloud-based solutions, and prioritize tasks based on resource availability.
- **Stakeholder Risks**
  - **Misalignment of Expectations:** Differences in what stakeholders expect versus what the model delivers.
  - **Preventative Steps:** Regular stakeholder meetings, clear communication of progress, and alignment on goals.

2. **Change management**

- **Anticipated Change:**

- **Scope Adjustments:** As the project progresses, the scope might need to be adjusted based on the findings.

- **Planned Response:** Establish a change request process, evaluate the impact of changes, and adjust timelines as needed.

- **Unexpected Changes:**

  - Communication Strategies: Ensure that unexpected changes are communicated promptly to all stakeholders and documented.

3. **End user involvement**

   - user centered design: engage with end users throughout the project using feedback to ensure that the product meets their needs and expectations

   - documentation: Create clear and comprehensive documentation that end-users can refer to, including a user guide.

Project schedule

Task	**Duration**
Project initiation	2 days
Data collection and preprocessing	2 weeks
Model development	1 week
Model validation and testing	1 week
Integration and deployment	1 week
User guide and documentation	1 weeks
Project closure	1 week

- **Contingency planning**
    - Buffer time: Allocate extra time for critical tasks where delays are most likely. I will add a buffer to model development and deployment
- **Resource availability**
    - **Tools and Technologies:** Ensure that the required software, tools, and technologies are available and accessible.

Alternative criteria for non- cost project

- **Time spent**
    - Estimate the total time spent on different phases of the project:
    - Data collection and preprocessing :1week
    - Model training: 1 week
    - Model evaluation: 1 week
    - Model integration and deployment: 2 weeks
- **Resource utilization**
    - Software usage: python, pandas, NumPy, matplotlib, seaborn, plotly, streamlit

Requirements Analysis

- **Use case: customer segmentation**
- **Objective:** Segment customers into distinct groups based on purchasing behavior for targeted marketing.

- **Primary actor:** sales team
- **Trigger:** the sales team needs to create customized marketing strategies for different customer segments.
- **Sequence of actions:**

- **Data Collection**: The system gathers data on customer demographics, purchasing behavior, and product usage.
- **Data Preprocessing**: Clean and prepare the data (handling missing values, feature engineering).
- **Model Training**: A classification algorithm (logistic regression, random forests, K-NN, support vector machine) is used to classify customers into segments.
- **Segmentation Analysis**: The system provides a detailed report on each segment's characteristics.
- **Action**: The sales team uses segmentation to tailor marketing strategies for each group.

System Design

- Top-Down System Design Overview

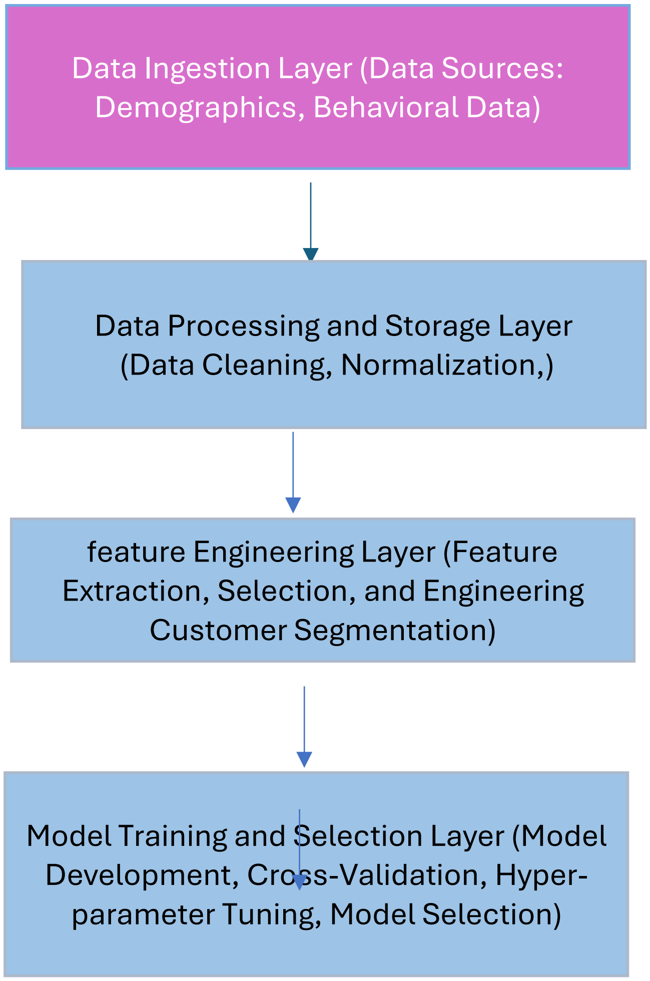
1. Data Ingestion Layer: handles the collection of raw data from various sources like customer interactions, sales records, demographic information.
    - Reason for inclusion: effective data collection is the foundation of any predictive model. Without accurate and comprehensive data, the model will not perform well.

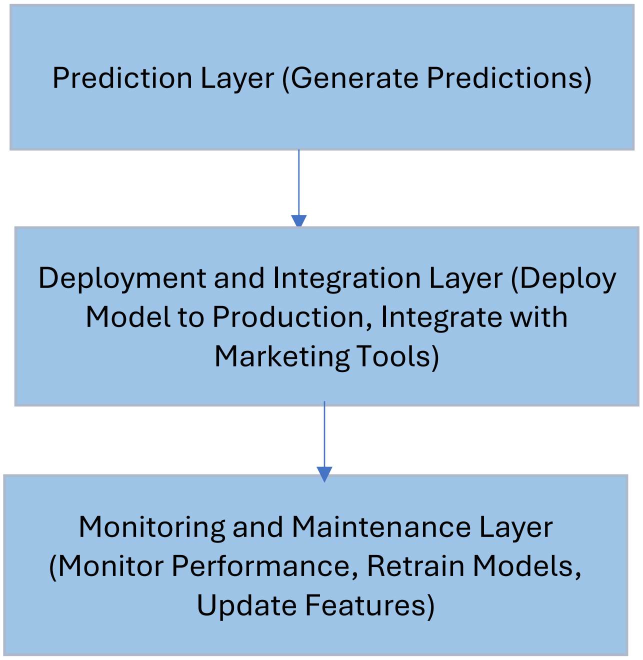
2. Data Processing and Storage Layer: after data is ingested, it undergoes processing such as cleaning (removing noise, handling missing values), normalization.
    - Reason for inclusion: Processed and well-organized data is crucial for accurate feature extraction and model training.
3. Feature Engineering Layer: this layer involves the transformation of raw data into meaningful features that can be used by machine learning models.
    - Reason for inclusion: the quality of features directly impacts the model's predictive performance. Proper feature engineering ensures that the model captures the underlying patterns in customer behavior.
4. Model Training and Selection Layer: In this component, various machine learning algorithms are trained on the processed data. Techniques such as cross-validation and hyperparameter tuning are used to select the best model.
    - Reason for inclusion: Different models may perform better on different datasets. This layer ensures that the most accurate and robust model is selected for deployment.
5. Prediction Layer: Once the model is trained, it can be used to generate predictions on new data.
    - Reason for inclusion: the output of this layer is actionable insights that can be used by marketing and sales teams to make data-driven decisions.
6. Deployment and Integration Layer: the selected model is deployed into a production environment and integrated with existing systems

- Reason for inclusion: deployment allows the model to be used in real-time applications, directly influencing business strategies. Integration ensures that the insights generated by the model are easily accessible to stakeholders.

7. Monitoring and Maintenance Layer: this component involves ongoing monitoring of the model's performance.

   - Reason for inclusion: machine learning models can degrade over time due to changing customer behavior. Regular monitoring and maintenance are necessary to ensure the model remains accurate and reliable.
   - System Design Diagram


Data Ingestion Layer (Data Sources: Demographics, Behavioral Data)
Data Processing and Storage Layer (Data Cleaning, Normalization,)
feature Engineering Layer (Feature Extraction, Selection, and Engineering Customer Segmentation)
Model Training and Selection Layer (Model Development, Cross-Validation, Hyper-parameter Tuning, Model Selection)


Prediction Layer (Generate Predictions)
Deployment and Integration Layer (Deploy Model to Production, Integrate with Marketing Tools)
Monitoring and Maintenance Layer (Monitor Performance, Retrain Models, Update Features)

Technical requirements

- **Software requirements**
  - **Programming Languages**: Python is recommended for its extensive libraries in data science (e.g., pandas, scikit-learn) and ease of integration with other systems.
  - **Machine learning libraries**: scikit-learn
  - **Visualization tools**: **Matplotlib/Seaborn**: For creating visualizations to analyze data and model outputs.
- **Hardware requirements**
  - **Server:** High-performance servers with multiple CPUs and GPUs to handle the computational load, streamlit cloud

Data science model

- **Pipeline stages:**

1. Data collection
2. Data processing
3. Feature engineering
4. Model selection
5. Model training
6. Model evaluation
7. Model deployment

Data collection:

- **Source of data:** I will use data from credible source which Kaggle repository
- **Data type:** structured data which is csv files containing customer demographic and behavioral information. The attributes are age, gender, graduated, marital status, profession, work experience, spending score, family size and segmentation

Data preprocessing:

- **Data cleaning:** handling missing values, outliers, and inconsistencies in data.
- **Data transformation:** standardization of features, converting categorical variables to numerical using techniques like ordinal and label encoding.

Feature engineering

- **Feature selection:** Identifying the most relevant features for the model to improve performance and reduce dimensionality.

Model selection

- **Algorithm Selection**: Choosing appropriate machine learning algorithms based on the problem type which classification (logistic regression, random forests, support vector machine, K-NN, decision tree)
- **Hyperparameter Tuning**: Using methods like grid search or random search to find the best hyperparameters for the chosen algorithms.

Model training

- **Split the data:** split the data into 80% training and 20% testing set
- **Training:** using the preprocessed data to train the selected models

  Cross-Validation: Validating the model's performance using techniques like k-fold cross-validation to ensure it generalizes well to unseen data.

Model evaluation

- **Performance Metrics**: Evaluating the models using metrics such as accuracy, precision, recall, F1-score for classification models
- **Model Comparison**: Comparing different models to select the best-performing one.

Model deployment

- **Deployment**: Integrating the model into the production environment where it can make real-time predictions.

Requirement Analysis: Reports

- **Types of reports**

1. **Data exploration report:** to summarize the initial dataset, highlight key statistics, and identify potential data quality issues (e.g., missing values, outliers).

- **Contents**
  - **Statistical data summary**: mean, median, standard deviation for numerical variables; frequency counts for categorical variables.
  - **Univariate analysis**: It aids in understanding the distribution of each variable, identifying patterns, outliers, and potential data quality issues
  - **Bivariate analysis:** helps in understanding the relationships between two variables, the input features and the target variable which is crucial for feature selection and engineering.

2. **Model performance report:** To present the performance of different machine learning models used for predicting customer behavior.

- **Contents**
  - **Model Metrics**: Accuracy, precision, recall, F1-score,
  - **Confusion Matrix**: Visualization of true positives, false positives, true negatives, and false negatives.
  - **Hyperparameter Tuning**: Results of hyperparameter optimization.

3. **Customer Segmentation Report:** To provide insights into the different customer segments

- **Contents**
  - **Segmentation Summary**: Overview of each segment (size, key characteristics).
  - **Segment Profiles**: Detailed descriptions of customer behavior within each segment.

Requirement Analysis: Screen definitions and layouts

For graphical user interface (GUI), I am using streamlit, it is an open-source framework for creating interactive web applications in Python. It is well-suited to building and custom web applications for machine learning models and data visualization. I will implement data preprocessing, exploratory data analysis, model development and evaluation. The app should allow users to access some of the data, its shape, its statistical summary, visual representation of the data, model performance with different metrics.

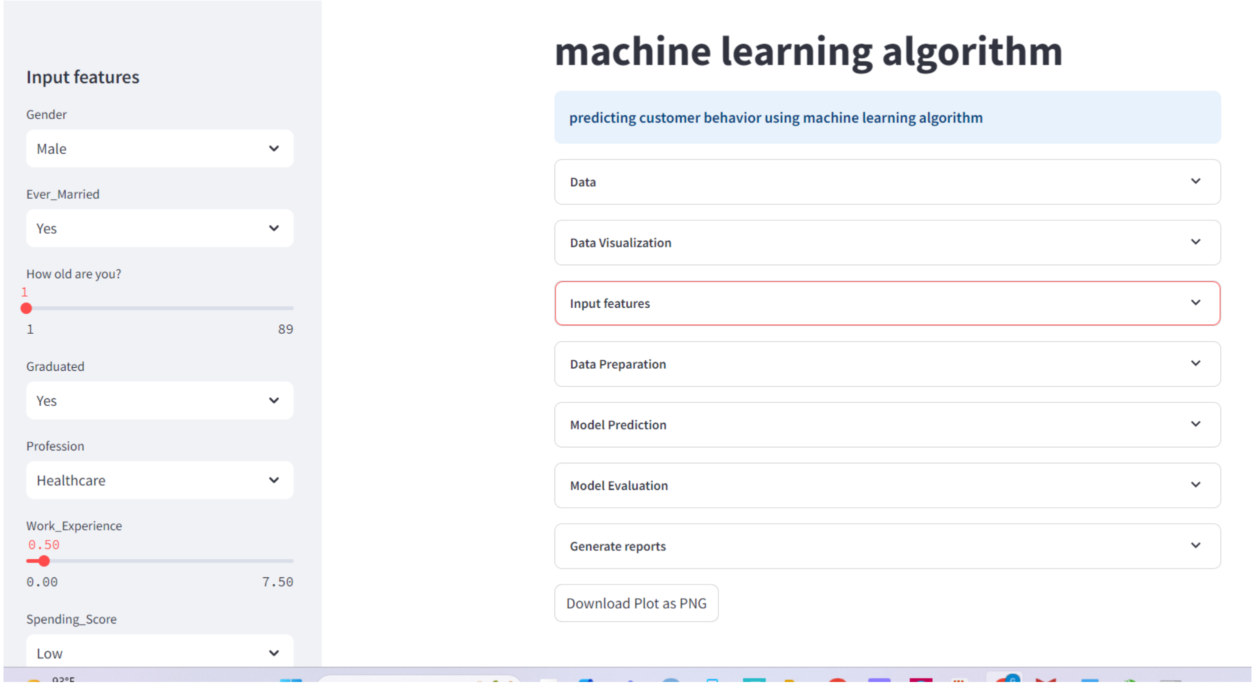
- **Home screen:** a brief introduction to the application and its purpose.
- **Data Exploration Screen**: allows users to explore and visualize the dataset.
- **Data visualization screen** allows users to explore and visualize using statistical summary, univariate analysis, bivariate analysis of features

- **Input feature screen** allows users to manipulate the input features using sliders, dropdowns to see the predicted segment accordingly.

- **Model prediction screen** allows users to see the predicted segment with its probabilities based on the input feature set by the user

- **Model evaluation screen** allows users to evaluate the performance of the model using accuracy and confusion matrix based on users' selection from the select box

- **Report screen** allows users to download reports generated by the system



Input features

Gender: Male

Ever_Married: Yes

How old are you?: 1 (1–89)

Graduated: Yes

Profession: Healthcare

Work_Experience: 0.50 (0.00–7.50)

Spending_Score: Low

machine learning algorithm

predicting customer behavior using machine learning algorithm

Security

The dataset that I am using for my capstone project is an automobile company dataset from Kaggle repository acquired from Analytics Vidhya. The dataset is under the license of CCO: public domain which means the creator has voluntarily waived all their copyright and related rights in the work, effectively placing it into the public domain. This means that:
No Copyright Restrictions: The work is free of copyright restrictions. Anyone can use, modify, distribute, and build upon the work without seeking permission from the creator
Free for All Uses: The work can be used for any purpose, including commercial and non-commercial uses, without any restrictions.
No Attribution Required: Unlike other Creative Commons licenses, a work under the CC0 license does not require users to credit the original creator, although doing so is often appreciated.
Research and Education: CC0 works can be freely used in research, educational materials, textbooks, and presentations without concerns about copyright infringement.

1. **Ethical data usage**
    - **Data Source Evaluation**: I prioritized using ethically sourced data, ensuring that all datasets employed in my project are publicly available with appropriate licenses
    - **Avoiding Harmful Outcomes**: I assessed the potential impact of the project's outcomes, particularly how insights derived from the data could be used. I tried to avoid bias by choosing the proper algorithms and data preprocessing methods to mitigate bias and ensure fairness.

2. **Privacy Considerations**
    - **Data Anonymization:** the creator of the dataset already implemented strong anonymization techniques to protect individuals' privacy. This includes removing or masking personally identifiable information (PII) to ensure that the data cannot be traced back to specific individuals.
    - **Transparency and Consent:** Although my project used publicly available datasets, I considered the importance of transparency and informed consent in data collection.
3. **Intellectual Property Rights (IPR)**

- **Respecting Ownership:** I ensured that all third-party content, including datasets, algorithms, and code, was used in accordance with their respective licenses.
- **Original Contributions:** I focused on contributing original work and avoided infringing on the intellectual property of others.

2. Milestone 2: Model Pipeline Design

Design Planning Summary

Customer behavior analysis helps business to better understand their customer needs, preferences and buying pattern, thus helps marketing team to tailor their efforts to reach out customers in the most efficient way for a better customer satisfaction and marketing strategies. Machine learning and predictive analysis are the most required tools for business to uncover and understand customer needs from huge collection of data. The objective of this project is to provide in-depth predictive analysis of customer behavior using machine learning algorithms which plays a vital role in decision making, improve profit rates of business, increase customer satisfaction, and reduce risk by identifying them at the early stage. By utilizing the existing literature, and machine learning algorithms I will explore and identify the factors that affect the experience and satisfaction of customers within the business. Thus, will enhance the business in many aspects such as:

- Personalized marketing: the product will help the business to create highly targeted marketing campaigns focused on the interest and behaviors of specific customer groups, leading to a more satisfying customer experience

- Efficient resource allocation: business can focus their marketing, sales, and product development effort on segments that respond positively which maximize revenue Increase
- customer retention: by meeting the specific needs of each segment business can build relationships with customers, increasing loyalty and reducing churn

With the increasing availability of customer data from various sources such as online transactions, social media, and customer service interactions, businesses are overwhelmed by the sheer volume of information. Traditional analysis methods are insufficient for extracting meaningful insights from these large datasets. As a result, companies struggle to identify key behavioral patterns and trends that could inform their strategies. The inability to harness this data effectively leads to missed opportunities and inefficiencies in targeting and customer engagement efforts.

The goal of the project is to predict customer behavior within the business using machine learning algorithms.   The project tries to answer how customer behavior is highly related to customer experience within the business. I will collect data from Kaggle repository. Once I acquire the required data, I will examine the data structure and its quality, checking for missing value, outliers, any inconsistences within the data set.   I will perform exploratory data analysis to understand the correlations and pattern of the variables. I will apply random forest, logistic regression, decision tree, K-NN and support vector machine and evaluate them using performance metrics to select the best fitting model for the dataset.

Overview of Design Concepts

The overall objective is to create a comprehensive system that efficiently processes customer data, generates predictive insights, and supports data-driven decision-making. This high-level design includes data acquisition, preprocessing, exploration, visualization, modeling, and result interpretation.

1. **Data Acquisition:** The first step in the model pipeline involves obtaining relevant customer data.
    - **Data source**: The data set I am using is from Kaggle repository acquired from Analytics Vidhya.
    - **Data format**: The dataset has two csv files train and test. the train set has 8068 rows and 11 columns, and the test set has 2627 rows and 11 columns
    - **Data volume**: The dataset has two csv files, train and test. The train dataset has (8068 ,11) and test dataset (2627,11). The dataset contains gender of the customer, marital status, age, is the customer graduate, profession of the customer, work experience, spending score and the target variable, customer segment.
2. **Data Preparation:** Once the data is collected, it undergoes a thorough cleaning and preprocessing phase to prepare it for analysis. This includes:

- **Data cleaning:** using isnull (), I check the null values replace them using mod for categorical columns and mean for numerical ones. Identifying and handling outliers effectively can prevent biasing the model, reducing its performance and hindering its interpretability. I use statical methods, interquartile range (IQR) to handle outliers in my data

- **Data transformation:** I transform the data into a required format using label and ordinal encoder, both the input features and the target variable
- **Feature engineering:** transforming raw data into features that are suitable for machine learning models and improves the performance of the model. I will apply normalization, scaling and encoding.

  Data splitting: splitting data into training and testing which 80 for training and 20% for testing then fitting models to the training data.

3. **Exploratory Data Analysis (EDA):**

   - **Data exploration:** Inspect data and compute descriptive statistics: using info (), describe (), value counts () to get insights on the descriptive overview of the data. to get a descriptive overview of the dataset by examining the mean, standard deviation, minimum and maximum value of the numerical feature. This helps to understand the overall structure and format of the dataset
   - **Visualization:** identify significant trends and patterns using matplotlib, seaborn and plotly. I use both univariate and bivariate analysis.

4. **Model selection**
   - **Objective:** develop a predictive model to predict the right group/segment of the new customers
   - **Algorithms:** choosing appropriate algorithms, classification from sklearn import logistic regression, K-NN, decision tree, random forest, and SVM.
   - **Rationale:**

     - **Logistic Regression**:
       Strengths:
     - Easy to implement and interpret, especially for binary classification.
     - Computationally efficient, works well with small to medium-sized datasets.
     - Provides probability estimates, which can be useful for decision-making.
       Weakness: - Assumes a linear relationship between the features and the log-odds of the outcome, which may not always be the case.
     - Struggles with complex relationships and interactions between features.

- Sensitive to Outliers**:** Outliers can influence the decision boundary significantly.

- **K-Nearest Neighbors (K-NN)**

- **Strengths:** Non-parametric**:** Makes no assumptions about the underlying data distribution.

- simple to understand**:** Intuitive and easy to interpret, especially in lower dimensions.

- Versatile**:** Can be used for classification, regression, and even density estimation.

- **Weakness:** Computationally Expensive**:** Becomes slow with large datasets, especially during the prediction phase.

- Memory Intensive**:** Requires storing the entire dataset, which can be impractical for large datasets.

- Sensitive to Feature Scaling**:** Performance is heavily dependent on the scale of features, requiring careful preprocessing

- Curse of Dimensionality**:** Performance deteriorates as the number of dimensions increases, as distances become less meaningful.

- **Decision Tree**

- **Strengths: i**nterpretability: easy to visualize and interpret, with clear decision rules.

- Non-parametric**:** No assumptions about the data distribution.

- Handles Non-linear Relationships**:** Can model complex interactions and non-linear relationships between features.

- Handles Both Numerical and Categorical Data: Versatile in handling different types of data.

- **Weaknesses:** Overfitting: Prone to overfitting, especially with deep trees and without pruning.

- Instability: Small changes in the data can result in a completely different tree structure.

- Biased with Imbalanced Data: Can be biased towards classes that dominate the training data.

- **Random Forest**

- **Strengths:** Reduced Overfitting**:** Combines multiple decision trees, which helps in reducing overfitting and improving generalization.

- Robust to Noise**:** The aggregation of multiple trees makes the model more robust to noise in the data.

- Feature Importance: Provides a measure of feature importance, aiding in feature selection.

- Handles Missing Data**:** Can handle missing data relatively well using surrogate splits.

- **Weaknesses:** Computational Complexity: Requires more computational resources and memory than a single decision tree, especially with a large number of trees.

- Interpretability: Although more accurate, it is less interpretable than a single decision tree.

- Slower Predictions**:** Prediction can be slower compared to simpler models, especially with a large ensemble of trees.

- **Support Vector Machine (SVM)**

- **Strengths:** Effective in High Dimensions: Works well with a large number of features, particularly when the number of dimensions exceeds the number of samples.

- Robust to Overfitting: Uses regularization (via the margin) to control overfitting, especially with the use of the kernel trick.

- Flexible**:** Can model complex relationships through different kernel functions (linear, polynomial, RBF, etc.).

- **Weaknesses:**

- Computationally Intensive: Training can be slow, especially with large datasets or complex kernels.

- Choice of Kernel**:** The performance of SVM depends heavily on the choice of the kernel and its parameters, which can be difficult to tune.

- Memory Usage**:** Requires substantial memory, especially for large datasets

- Interpretability: The decision boundary is less interpretable compared to simpler models like logistic regression or decision trees.

5. **Model training and Evaluation**

- **Training process**: Split the data into 80% training and 20% testing then train the machine learning model.
- **Evaluation metrics**: evaluate model performance using classification metrics such as accuracy, F1 score and confusion matrix
  - Classification accuracy: is a fundamental metric for evaluating the performance of a classification model, providing a quick snapshot of how well the model is performing in terms of correct predictions.

  - F1 score: it is a harmonic mean between recall and precision, its range is [0,1]. This metric tells us how precise and robust our model is
  - Precision: it is a measure of a model's performance and informs us how many of the positive predictions made by the model are actually correct
  - confusion matrix is a matrix that summarizes the performance of the model and displays the number of accurate and inaccurate instances based on the model's prediction.

6. **Result Interpretation and Business Insights**

- **Model interpretability:** Interpret which model from the five models performs well based on the performance metrics.

- **Business actionability:** provide actionable insights to give recommendations for the sales team based off the data set in a way that non-technical audience could understand. For each Segment (A, B, C, D): interpret professional inclination, dominant age group, family size, marital status percentage, educational level and spending score. Thus, will help the sales team to create highly targeted marketing campaigns focused on the interest and behaviors of specific customer groups, allocate resources effectively, and increase loyalty and reduce churn.

Detailed Model pipeline Design: Overview

- **The goal of the model**: to develop a predictive model that can classify customers into segments based on their behavior, allowing for more targeted marketing and personalized services.
- **Overview of the Machine Learning Pipeline:** the pipeline will start with raw customer data from Kaggle repository, proceed through preprocessing and cleaning, followed by model training, evaluation, and finally generating predictive insights that can directly impact marketing and sales strategies.

The following sections present how each stage of the model pipeline aligns with the business case.

Stages of the model pipeline

- **Stage 1: Data collection and ingestion**
- **Business relevance: t**he first step in predicting customer behavior is gathering relevant data from various sources, in my case from Kaggle repository. This data serves as the foundation for training the machine learning models.
- **Impact**: Timely and accurate data collection ensures that the model remains updated with the latest customer behavior trends, allowing businesses to adapt their strategies based on near-real-time information.
- **Stage2: Data Preprocessing and Cleaning**
- **Business relevance:** Data preprocessing is essential for preparing raw data for analysis. Data from different sources may have missing values, inconsistencies, or outliers that need to be addressed before it can be used to train models.
- **process**: This stage includes handling missing values, normalizing and scaling data, and feature engineering to enrich the dataset. It ensures that the data is consistent and clean, which in turn leads to more accurate predictions
- **Impact**: Proper data cleaning prevents the model from making erroneous predictions based on noisy or incomplete data, ensuring that business decisions are based on reliable insights.
- **Stage 3: Exploratory Data Analysis (EDA) and Feature Selection**
- **Business Relevance**: EDA helps identify key patterns, correlations, and trends within the data, which are critical for understanding customer behaviors.

- **process**: Visualization tools (e.g., Matplotlib, Seaborn) and statistical methods will be used to explore the data and select relevant features that have the most predictive power.
- **Integration**: The insights gained from EDA can be communicated to business stakeholders, allowing them to identify important customer segments and develop targeted strategies.
- **Impact**: EDA and feature selection improve the model's efficiency by reducing the dimensionality of the data and ensuring that the most important features are used in the model, ultimately leading to more precise and actionable insights for the business.

- **Stage 4: Model Selection and Training**

- **Business Relevance**: Model selection is crucial for finding the most effective machine learning algorithm to predict customer behavior. The choice of algorithm directly affects the accuracy and usefulness of the predictions.
- **Process**: multiple models will be trained and tested, including Random Forest, decision tree, logistic regression, support vector machine and K-NN. Hyperparameter tuning (via Grid Search) will be conducted to optimize model performance.
- **Impact**: By selecting the best model based on accuracy, precision, and recall, the business will have access to reliable predictions, improving decision-making related to marketing strategies, customer segmentation, and resource allocation.

- **Stage 5: Model Evaluation and Validation**

- **Business Relevance**: Model validation ensures that the selected model performs well not only on training data but also on unseen data. This is critical for building trust in the model's predictions.
- **Process**: metrics such as accuracy, F1-score, precision, and recall will be used to assess how well the model generalizes
- **Impact**: A well-validated model will produce more reliable predictions, reducing the risk of business decisions based on inaccurate forecasts. This helps mitigate financial risks associated with customer churn or misallocated resources.

- **Stage 6: Model Deployment**

- **Business Relevance**: Once the model has been trained and validated, it needs to be deployed into cloud for real-time predictions.
- **Process**: The model will be integrated into the streamlit cloud via cloud services for real-time predictions.
- **Impact**: The deployment of the model ensures continuous predictions that align with the company's day-to-day operations, enhancing efficiency and decision-making in real time.

- **Stage 7: Model Monitoring and Maintenance**
- **Business Relevance**: Customer behavior can change over time, and models may experience performance degradation as a result. Continuous monitoring and maintenance of the model are essential to ensure long-term success.
- **Process**: Model performance will be monitored using performance metrics and drift detection. Retraining mechanisms will be implemented to update the model with new data as necessary.
- **Impact**: By maintaining the model over time, the business ensures that predictions remain accurate and relevant to evolving customer behaviors, thereby safeguarding long-term business growth and customer retention.

Data sources

- The data set I am using is from Kaggle repository. The dataset has two csv files train and test. The train set has 8068 rows and 11 columns, and the test set has 2627 rows and 11 columns. The data fields are as follows:
- Data fields:

Variables	**Definitions**
Id.	Unique Id
Gender.	Gender of the customer
Ever married.	Marital status of the customer
Age	age of the customer
Graduated	Is the customer a graduate
Profession	profession of the customer
Work experience	work experience of the customer in years
Spending score.	Spending score of the customer
Family size	number of family members for customer (including the customer)
Var_1.	anonymized category for the customer
Segmentation	target variable (customer segments)

Dataset Types and Formatting

- Id is int64, its integer data type
- Gender is object, which has male and female binary type
- Ever _married is object. is a Boolean data type with yes/no type
- Age is int64, it ranges from 18 to 89 years old with mean of 43.5
- Graduated is object, a Boolean data type with yes/no answer

- Profession is object, it a categorical variable which includes different professional status of the customer such as healthcare, engineer, lawyer, entertainment, artist, executive, doctor, homemaker, marketing
- Work experience is float64, ranges from 0 to 14 years old with mean of 2.64
- Spending score is object, which includes low, average and high
- Family size is float64, ranges from 1 to 9 with mean of 2.64
- Var_1 is object, includes cat_1, cat_2, cat_3, cat_4, cat_5, cat_6, cat_7
- Segmentation is object, has A, B, C, D segments

Data cleaning procedures

Data cleaning is a critical step in ensuring that the datasets used for predicting customer behavior are accurate, consistent, and ready for analysis. The following procedures outline the comprehensive and systematic approach to cleaning the data, ensuring it is of high quality before feeding it into machine learning models.

- **Handling Missing Values**
  - Identify missing values using functions such as `isnull ()` in pandas
  - For numerical columns (age, family size, work experience), impute missing values using median imputation
  - For categorical columns (graduate, ever married, profession, category), use mode imputation
- **Justification**: Missing data can skew the analysis and predictions. Imputing with median maintains the distribution for numerical data, while mode imputation ensures that categorical variables remain consistent.
- **Removing Duplicates**
  - Check for duplicate records using drop_duplicates `()` in pandas.
  - Duplicate records can introduce bias and incorrect results. Ensuring that each record represents a unique event or entity prevents data redundancy and maintains dataset integrity.
- **Handling Outliers**
  - **Procedure**: Use statistical methods such as the Interquartile Range (IQR) to detect outliers in numerical columns (age, work, experience, family size)
  - **Justification**: Outliers can distort model predictions, especially when using algorithms sensitive to extreme values. Removing erroneous outliers improves model accuracy.
- **Dealing with Inconsistent Data**
  - identify inconsistencies in categorical variables (e.g., different spellings for the same value, such as “Male”, “male”, and “M”).
  - **Justification**: Inconsistent data leads to fragmentation and poor model performance. Standardizing categorical entries ensures the data is clean and uniform across datasets, preventing issues like duplicate categories.
- **Removing Irrelevant Data**

- identify and remove irrelevant columns that do not contribute to customer behavior prediction (e.g. ID)
- Ensure all remaining fields are directly related to customer behavior analysis
- **Justification**: Irrelevant data can introduce noise into the analysis and slow down the processing. Removing unnecessary columns streamlines the datasets and improves computational efficiency.

- **Data Encoding for Categorical Variables**

- Convert categorical data (segmentation, profession, category) into numerical format using label encoding
- Convert categorical columns (spending score) into numerical format using ordinal encoding
- Convert categorical columns (gender, graduated, ever married) into numerical format using predefined mapping
- **Justification**: Machine learning models typically require numerical inputs. Encoding categorical variables allows models to process them correctly, ensuring that important categorical data is not lost or misrepresented.

- **Normalization/Standardization of Numerical Data**

- normalize numerical features using Min-Max scaling or Z-score normalization
- Standardize data where necessary to ensure that variables are on a similar scale
- **Justification**: Normalization ensures that features with different scales do not disproportionately influence the model.

Data Exploration

Data exploration is a critical phase in understanding the structure, patterns, and relationships within the data before applying machine learning algorithms. This phase includes both **descriptive statistics** and **data visualization** methods to uncover insights and guide further processing.

- **Summary statistics**
  Descriptive Statistics:

- Use functions such as describe () in pandas to get key summary statistics (mean, median, standard deviation, min, max) for numerical variables like Age, family size and work experience

- For categorical variables (gender, marital status, profession, graduated), use value_counts () to assess the distribution of each category.

- Summary statistics provide an immediate overview of the data's distribution and spread. By identifying key metrics such as the mean and variance, we can better understand how customers' behavior varies across different segments.

- **Univariate Analysis:** univariate analysis involves looking at the distribution of a single variable and is excellent way to understand a dataset's range and spread

- **Histograms and KDE Plots**: Use histograms and Kernel Density Estimation (KDE) plots to visualize the distribution of continuous variables such as age, family size and work experience
- **Bar Plots for Categorical Variables**: Use bar plots to show the frequency distribution of categorical variables
- **Bivariate Analysis:** looks at the relationship between two variables using pair plots, grouped bar charts, histogram and thus gives us a better understanding of how two variables interact.
- **Scatter Plots**: Use scatter plots to explore relationships between two numerical variables, such as age and work experience . Add trend lines to examine the correlation
- **Box Plots and Violin Plots**: Use these plots to explore the distribution of numerical variables across different categories (age across different spending score). Violin plots, in particular, provide detailed insights into both the distribution and density
- **Correlation Matrix**: Visualize the correlation between numerical variables using a heatmap of the correlation matrix.

- **Target variable analysis:** understanding the distribution of the target variable is crucial for model performance.

- **Class Imbalance Check**: Visualize the distribution of the target variable (Segmentation) to check for class imbalance. Use pie charts to assess how balanced the target variable is.

Data model

- **Data objects:** the primary data objects in this project are designed to capture customer demographics, behavior and product details. Below is a detailed breakdown of each object:

- **Attributes:**
- Customer_ID : Unique identifier for each customer
- Gender: Male/Female
- Age: Customer's age in years
- Family_Size: Number of family members
- Ever married: marital status of the customer
- Gradated: graduate status of the customer
- Profession: professional status of the customer
- Work experience: work experience of the customer
- Spending score: spending score of the customer
- Var_1: anonymized category
- Segmentation: target variable

Here is the detailed information of the data set using info () from pandas :

- Id is int64, it integer data type
- Gender is object, which has male and female binary type
- Ever married is object. is a Boolean data type with yes/no type
- Age is int64, it ranges from 18 to 89 years old with mean of 43.5
- Graduated is object, a Boolean data type with yes/no answer
- Profession is object, it a categorical variable which includes different professional status of the customer such as healthcare, engineer, lawyer, entertainment, artist, executive, doctor, homemaker, marketing
- Work experience is float64, ranges from 0 to 14 years old with mean of 2.64
- Spending score is object, which includes low, average and high
- Family size is float64, ranges from 1 to 9 with mean of 2.64
- Var_1 is object, includes cat_1, cat_2, cat_3, cat_4, cat_5, cat_6, cat_7
- Segmentation is object, has A, B, C, D segments

Methodology

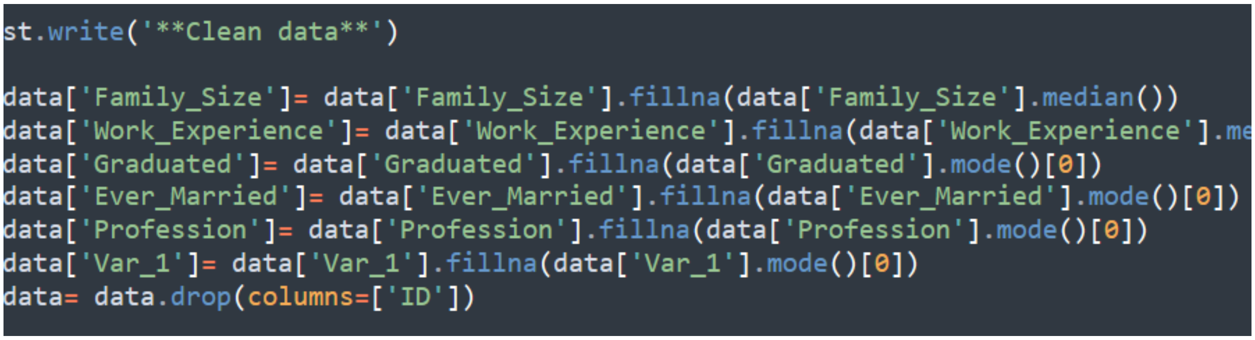
The methodology for predicting customer behavior using machine learning is structured around several core steps: **data preprocessing**, **feature engineering**, **model selection**, **model training**, **evaluation**, and **interpretation of results**. This approach ensures that the data is correctly processed, the models are applied appropriately, and the results are interpreted in a way that drives actionable insights.

1- Data Preprocessing methodology: Ensure the data is clean, consistent, and ready for modeling.

- **Handling Missing Data**:

- Identify missing values in input features such as Age, family size, work experience, spending score, etc. using isnull () and address them based on the nature of the missing data.

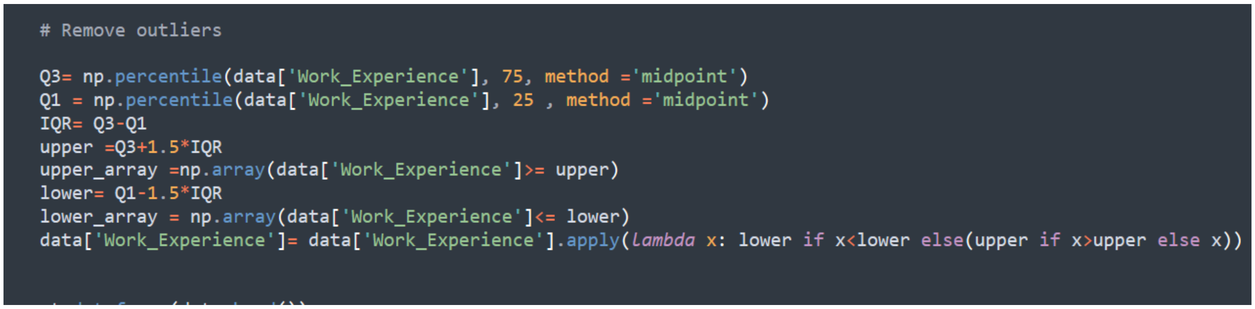
-



```
st.write('**Clean data**')

data['Family_Size']= data['Family_Size'].fillna(data['Family_Size'].median())
data['Work_Experience']= data['Work_Experience'].fillna(data['Work_Experience'].me
data['Graduated']= data['Graduated'].fillna(data['Graduated'].mode()[0])
data['Ever_Married']= data['Ever_Married'].fillna(data['Ever_Married'].mode()[0])
data['Profession']= data['Profession'].fillna(data['Profession'].mode()[0])
data['Var_1']= data['Var_1'].fillna(data['Var_1'].mode()[0])
data= data.drop(columns=['ID'])
```

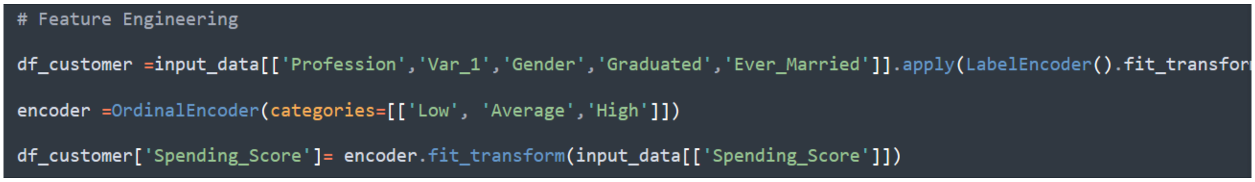
- **Justification**: Missing data, if not handled properly, and negatively impact model performance. Methods like median imputation for numerical features or mode imputation for categorical features ensure that missing data is appropriately filled.
- **Outlier Detection and Treatment**
- Outliers will be detected using methods like interquartile ranges (IQR), or visual techniques like box plots.
- Depending on the nature of the outliers, they may be transformed, capped (clipping), or removed if they are determined to be errors.



```
# Remove outliers

Q3= np.percentile(data['Work_Experience'], 75, method ='midpoint')
Q1 = np.percentile(data['Work_Experience'], 25 , method ='midpoint')
IQR= Q3-Q1
upper =Q3+1.5*IQR
upper_array =np.array(data['Work_Experience']>= upper)
lower= Q1-1.5*IQR
lower_array = np.array(data['Work_Experience']<= lower)
data['Work_Experience']= data['Work_Experience'].apply(Lambda x: lower if x<lower else(upper if x>upper else x))
```

- **Justification**: Outliers can skew the model's performance, especially in models sensitive to extreme values like K-nearest neighbors.
- **Data Transformation**
- **Normalization or standardization**
- For algorithms sensitive to scaling (e.g., KNN, Support Vector Machines), continuous features will be standardized (mean=0, standard deviation=1) or normalized (range from 0 to 1).
- **Categorical Encoding**: Categorical variables will be encoded using methods like one-hot encoding for nominal data and label encoding or ordinal encoding for ordinal variables.



```
# Feature Engineering

df_customer =input_data[['Profession','Var_1','Gender','Graduated','Ever_Married']].apply(LabelEncoder().fit_transfor

encoder =OrdinalEncoder(categories=[['Low', 'Average','High']])

df_customer['Spending_Score']= encoder.fit_transform(input_data[['Spending_Score']])
```

-
- **Justification**: Normalization/standardization ensures that all features contribute equally to the model, while encoding ensures that categorical data is handled appropriately for machine learning algorithms.

2- Data Analysis Methodology: This phase involves selecting appropriate machine learning algorithms and evaluating their performance.

- **Model selection:** the choice of models will depend on the target outcome (regarding my project it is classification). Common models include
- **Logistic regression, random forest, decision tree, support vector machine, K-NN**

- **Model evaluation: Accuracy, Precision, Recall, F1 Score**: These metrics help evaluate how well the model classifies customers.

3- interpreting the analysis results

- **Model interpretability:** essential, especially in business contexts where stakeholders need to trust and understand the model's decision-making process.
- **Error analysis:** error analysis helps in refining the model by identifying patterns in errors, indicating whether the model needs more training data, feature engineering, or a different algorithm.
- **Confusion Matrix**: For classification tasks, a confusion matrix will highlight where the model is making incorrect predictions (false positives/false negatives).
- **Actionable Insights:** will provide business teams with actionable insights for customer segmentation and personalized marketing strategies.

- **Infrastructure and Environment Configuration:**

- before building and deploying the machine learning models, it's crucial to configure the environment to support the development and implementation of the project.
- **Local Development Setup**: Install relevant libraries like scikit-learn, pandas, NumPy, matplotlib, seaborn, streamlit
- set up the Python version (e.g., Python 3.8 or higher) to ensure compatibility with libraries.
- Ensuring the environment is set up correctly reduces the risk of version conflicts or missing dependencies, facilitating smooth model development.
- **Cloud set up:** sign up for streamlit cloud for deploying streamlit app, making it accessible via a web interface
- Create GitHub repository for an easy connection with streamlit community cloud

- **Security configuration**
- ensure that any personally identifiable information (PII) is anonymized before being used in the mode
- Protecting sensitive data ensures compliance with data protection regulations like GDPR and builds trust with stakeholders.
- set up role-based access controls (RBAC) to ensure that only authorized personnel can access the data or modify the model.
- **Model Deployment Configuration**
- Configure tools like Docker for containerization, ensuring that the model can be easily deployed across different environments (development, testing, production).

- **Security**

The dataset that I am using for my capstone project is an automobile company dataset from Kaggle repository acquired from Analytics Vidhya. The dataset is under the license of CCO: public domain which means the creator has voluntarily waived all their copyright and related rights in the work, effectively placing it into the public domain. This means that:
No Copyright Restrictions: The work is free of copyright restrictions. Anyone can use, modify, distribute, and build upon the work without seeking permission from the creator
Free for All Uses: The work can be used for any purpose, including commercial and non-commercial uses, without any restrictions.
No Attribution Required: Unlike other Creative Commons licenses, a work under the CC0 license does not require users to credit the original creator, although doing so is often appreciated.
Research and Education: CC0 works can be freely used in research, educational materials, textbooks, and presentations without concerns about copyright infringement.

4. **Ethical data usage**
   - **Data Source Evaluation**: I prioritized using ethically sourced data, ensuring that all datasets employed in my project are publicly available with appropriate licenses
   - **Avoiding Harmful Outcomes**: I assessed the potential impact of the project's outcomes, particularly how insights derived from the data could be used. I tried to avoid bias by choosing the proper algorithms and data preprocessing methods to mitigate bias and ensure fairness.

5. **Privacy Considerations**
   - **Data Anonymization:** the creator of the dataset already implemented strong anonymization techniques to protect individuals' privacy. This includes removing or masking personally identifiable information (PII) to ensure that the data cannot be traced back to specific individuals.
   - **Transparency and Consent:** Although my project used publicly available datasets, I considered the importance of transparency and informed consent in data collection.
6. **Intellectual Property Rights (IPR)**
   - **Respecting Ownership:** I ensured that all third-party content, including datasets, algorithms, and code, was used in accordance with their respective licenses.
   - **Original Contributions:** I focused on contributing original work and avoided infringing on the intellectual property of others.

3. Milestone 3: Implementation

1. System Entities

1.1 End- users: Individuals who interact with the app to input customer data, explore the dataset, visualize data, make predictions, evaluate the model, and generate reports.

1.2 Input features

- **Input form fields (located on the left side of the App)**
- **Gender:** Categorical input (Male, Female) in select box
- **Ever married:** Binary input indicating whether the customer is married (Yes/No) in a select box
- **Age:** Numeric input for the customer's age using slider from 18 to 89 years
- **Graduate: Binary input indicating whether the customer is a graduate** (Yes/No)
- **Profession:** Categorical input for customers profession in a select box(healthcare, engineer, doctor, lawyer, executive, marketing**)**
- **Work experience:** Numeric input for years of work experience using slider 0 to 7.5 years
- **Spending score:** categorical input representing the customer's spending score (low, Average, high) in select box
- **Family size:** Numeric input for the size of the customer's family in slider (1 to 9)
- **Var_1:** anonymized categorical input of the customer in select box (cat_1, cat_2, cat_3, cat_4, cat_5, cat_6, cat_7)

1.3 Data Exploration (Right side of the App):

- **Head of the Raw Data**: Displays the first 10 rows of the dataset to provide a preview.
- **Data Shape**: Shows the dimensions of the dataset (number of rows and columns).

- **Data Info**: Provides a summary of the dataset, including data types and non-null counts.
- **Data Types**: Lists the types of data (e.g., integer, float, categorical) for each column in the dataset.
- **Download CSV Button**: Allows users to download the dataset in CSV format

1.4 Data Visualization (Right Side of the App)

- **Data Summary**: Statistical summary of numeric features (e.g., mean, median, standard deviation).
- **Univariate Analysis:**
- **Histograms**: Visualizations for the distribution of features such as Age, Work Experience, and Family Size. Users can select which feature to visualize from a select box.
- **Value Counts**
- **Value Count Selection**: Allows users to select from features like Gender, Ever Married, Profession, Spending Score, and Graduated. The app generates the value count and displays it along with a bar graph of the selected feature.
- **Bivariate Analysis**:
- **Plotly Graphs**: Interactive Plotly visualizations that show relationships between the target variable (customer segmentation) and input features (Gender, Age, Ever Married, Profession, Spending Score, Graduated). Users select features from a dropdown menu, and the app generates bar graphs based on the selected input.

1.5 Data Preparation (Right Side of the App)

- **Feature engineering**
- **Encoded Input Features**: Displays a table showing the encoded form of input features using encoding methods such as Label Encoding and Ordinal Encoding
- **Encoded Target Variable**: Displays the encoded form of the target variable(segmentation)

1.6 Model Prediction (Right Side of the App)

- **Random Forest Model**: The machine learning model used to predict customer segments based on user input features.
- **Predicted segments:** Displays the probability of each segment (A, B, C, D) in a progress column.
- Shows the predicted customer segment with the highest probability in a single row.

1.7 Model Evaluation (**Right** side of the App**)**

- **Select matrix to display**: Allows users to choose between different evaluation metrics (e.g., accuracy or confusion matrix) for model assessment.
- **Accuracy Metrics**: Provides the accuracy score of the model to assess its performance.
- **Confusion Matrix**: Displays a confusion matrix to evaluate the classification model's performance.

1.8 Reports

- **Numerical Feature Insights**
- **Violin Plots**: Visualizes distributions of numerical features (Age, Work Experience, Family Size) against the target variable. Users select features from a dropdown menu.
- **Download and Print Reports**: Users can download or print reports with these visualizations.
- **Categorical Feature Insights**
- **Bar Graphs**: Visualizes distributions of categorical features (Profession, Spending Score, Graduated, Marital Status) against the target variable. Users select features from a dropdown menu.
- **Download and Print Reports**: Users can download or print reports with these visualizations.

1.9 System Interfaces

- **User Interface (UI)**: Layout featuring input fields on the left and functional sections (data exploration, visualization, preparation, prediction, evaluation, and reporting) on the right side.

2. Functional requirements:

2.1 Data ingestion:

- the system allows users to upload data in CVS format
- the system allows users to explore the data shape and information

2.2 data preprocessing:

- the system handles missing value, outliers, and inconsistencies
- the system allows users to check the cleaned data
- the system must perform featuring engineering, transforming features that are suitable for machine learning

2.3 data visualization:

- the system allows users to check the statistical data summary

- the system allows users to check and explore the distribution of a single variable (univariate analysis)
- it also allows users to explore and understand how the input feature interact with the target variable

2.4 model training and prediction: the system trains the random forest classifier using the input features

- the system makes predictions based on the input features
- the system display customer segments [A, B, C, D] with its probability in a progress column

2.5 model evaluation: the system allows users to evaluate the performance of the model using accuracy and confusion matrix

2.6 visualization:

- The system provides interactive charts (histogram, violin, bar charts) based on users' need.

2.7 Report generation: the system allows users to download plots with its reports

2.8 User interaction: the system provides easy user interaction via streamlit for users to interact with data, visualize it and make predictions.

2.9 Performance and Scalability ensures the app responds quickly to user inputs and predictions.

2.10 Security: ensure user input data is handled securely and comply with data privacy regulations (e.g., GDPR).

3. Implementation plan:

2.1 Libraries:

- Pandas for data manipulation and preprocessing.
- Scikit-learn for model development and evaluation.
- Matplotlib/Seaborn for data visualization.
- Streamlit for building an interactive front-end interface.
- NumPy for numerical operations.

2.2 Software:

I use Streamlit Cloud for deploying the Streamlit app, making it accessible via a web interface.

2.3 Strategy for implementation:

Phase1**:** building the core component of the product such as data processing pipeline, building model, prediction then testing the functionality in my local server
Phase2**:** push all the files to GitHub repository for an easy connection with streamlit community cloud
Phase **2:** deployment the product into the streamlit community cloud where end users can access the product

2.4 Potential impacts: checking streamlit community cloud CPU if it can handle heavy loads, and user interactions

2.5 Ensuring the product is available for end users: provide end users user guide on how to operate the system

4. Introduction

4.1 Overview of the project: Customer behavior analysis helps business to better understand their customer needs, preferences and buying pattern, thus helps marketing team to tailor their efforts to reach out customers in the most efficient way for a better customer satisfaction and marketing strategies. The objective of this project is to provide in-depth predictive analysis of customer behavior using machine learning algorithms which plays a vital role in decision making, improve profit rates of business, increase customer satisfaction, and reduce risk by identifying them at the early stage.

4.2 Scope: An automobile company has plans to enter new markets with their existing products and after intensive market research, they've realized that the behavior of the new market is like their existing market. In their existing market, the sales team has classified all customers into 4 segments (A, B, C, D). Then, they performed segmented outreach and communication for a different segment of customers. This strategy has worked exceptionally well for them. Accordingly, they plan to use the same strategy for the new markets and my data product is aimed to help the sales team to predict/classify the right group of the new customers (A, B, C,D) using machine learning algorithm. The data product will enhance sales, customer experience and overall marketing strategies. Key advantages of the data product:

- Personalized marketing: the product will help the sales team to create highly targeted marketing campaigns focused on the interest and behaviors of specific customer groups, leading to a more satisfying customer experience
- Efficient resource allocation: business can focus their marketing, sales, and

- product development effort on segments that respond positively which maximize revenue
- Increase customer retention: by meeting the specific needs of each segment business can build relationships with customers, increasing loyalty and reducing churn

5. **System requirements**

5.1 Software requirements: I use python, and libraries such as pandas, NumPy, matplotlib, plotly, seaborn, scikit- learn   streamlit

5.2 Installation instructions: for setting up the environment and installing dependencies I use pip command

6. **Dataset overview:**

6.1 Dataset description: The dataset is from Kaggle repository, it has two csv files train and test. The train set has 8068 rows and 11 columns, and the test set has 2627 rows and 11 columns.

6.2 Source of the data: the dataset is from Kaggle repository, and I did the necessary preprocessing such as

- handling missing values using median for categorical variables and mode for numerical variables
- removing outliers from the profession column using interquartile range (IQR)
- Feature engineering: transforming raw data into features that are suitable for machine learning models and improves the performance of the model using normalization, scaling and encoding

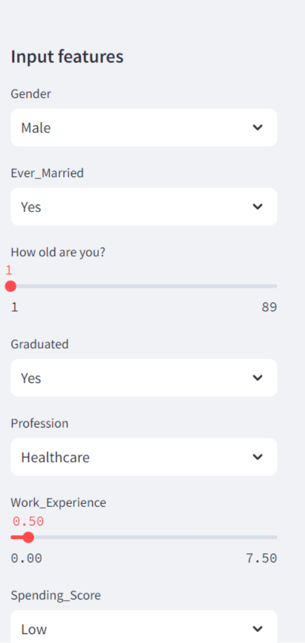
6.3 data fields:

Variables	**Definitions**
Id.	Unique Id
Gender.	Gender of the customer
Ever married.	Marital status of the customer
Age	Age of the customer
Graduated	Is the customer a graduate
Profession	profession of the customer

Work experience	work experience of the customer in years
Spending score.	Spending score of the customer
Family size	number of family members for customer (including the customer)
Var_1.	anonymized category for the customer
Segmentation	target variable (customer segments)

7. **step -by step instructions for using the product**

7.1 loading the data: load the data by clicking the download button



Input features

Gender

Male

Ever_Married

Yes

How old are you?

1

1 89

Graduated

Yes

Profession

Healthcare

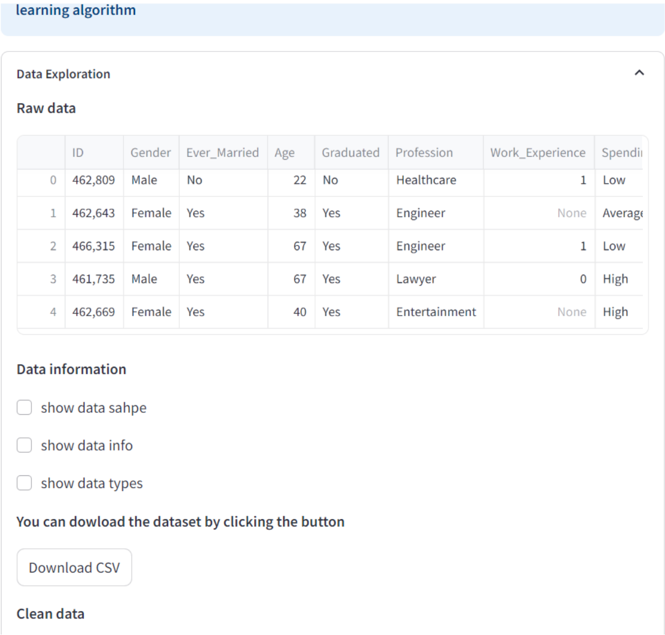
Work_Experience

0.50

0.00 7.50

Spending_Score

Low



learning algorithm

Data Exploration

Raw data

	ID	Gender	Ever_Married	Age	Graduated	Profession	Work_Experience	Spendi
0	462,809	Male	No	22	No	Healthcare	1	Low
1	462,643	Female	Yes	38	Yes	Engineer	None	Average
2	466,315	Female	Yes	67	Yes	Engineer	1	Low
3	461,735	Male	Yes	67	Yes	Lawyer	0	High
4	462,669	Female	Yes	40	Yes	Entertainment	None	High

Data information

☐ show data sahpe

☐ show data info

☐ show data types

You can dowload the dataset by clicking the button

Clean data

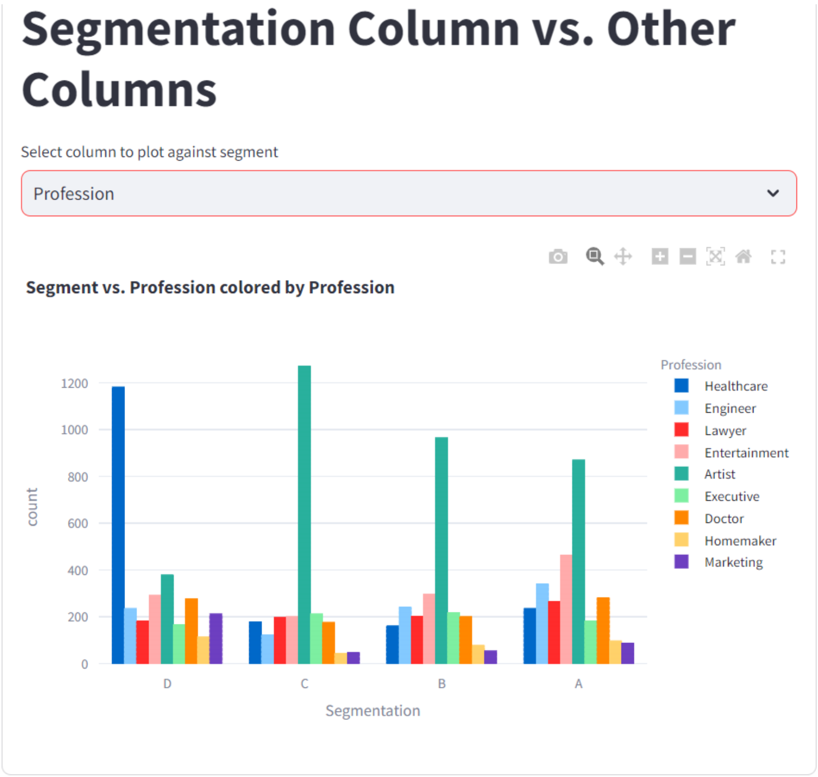
7.2 data exploration: check the shape and information of the data by using show data shape, show data info checkbox

7.3 data visualization: I am visualizing input features and the target variable, it includes

- statistical data summary,
- from the select box select a numerical column for histogram for univariate analysis of age, work experience and family size.

- Form the select box Select a categorical for value count of categorical features

- Bivariate analysis, from the select box select input features such as age, gender etc against the target variable segmentation to understand and explore their relationships



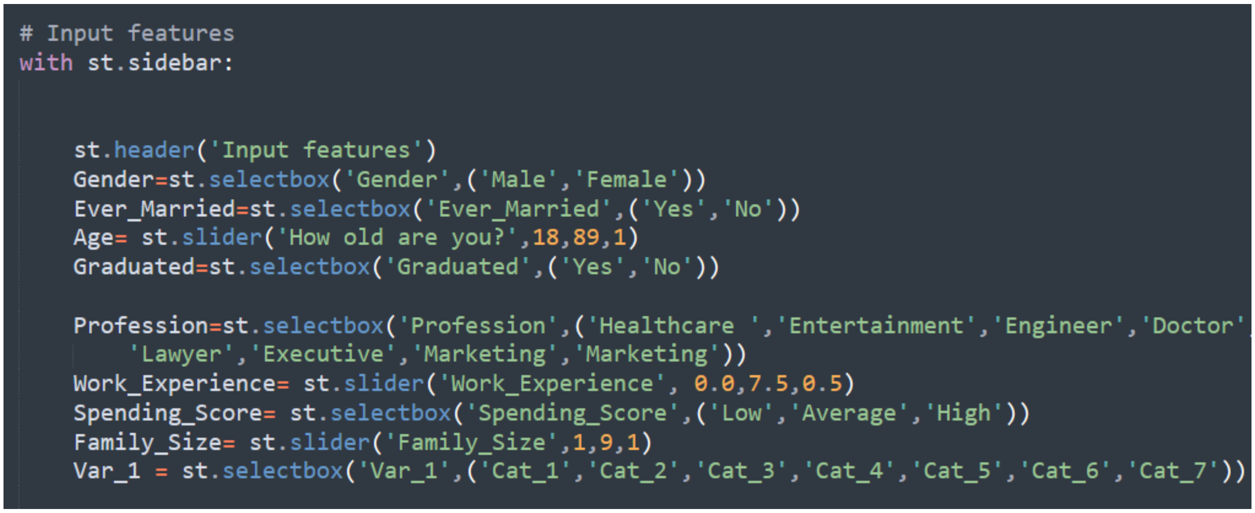
Segmentation Column vs. Other Columns

Select column to plot against segment

Profession

7.4 Input features: for clarity and simplicity I establish all input features at the sidebar using select box and slider so that users can easily manipulate as they want

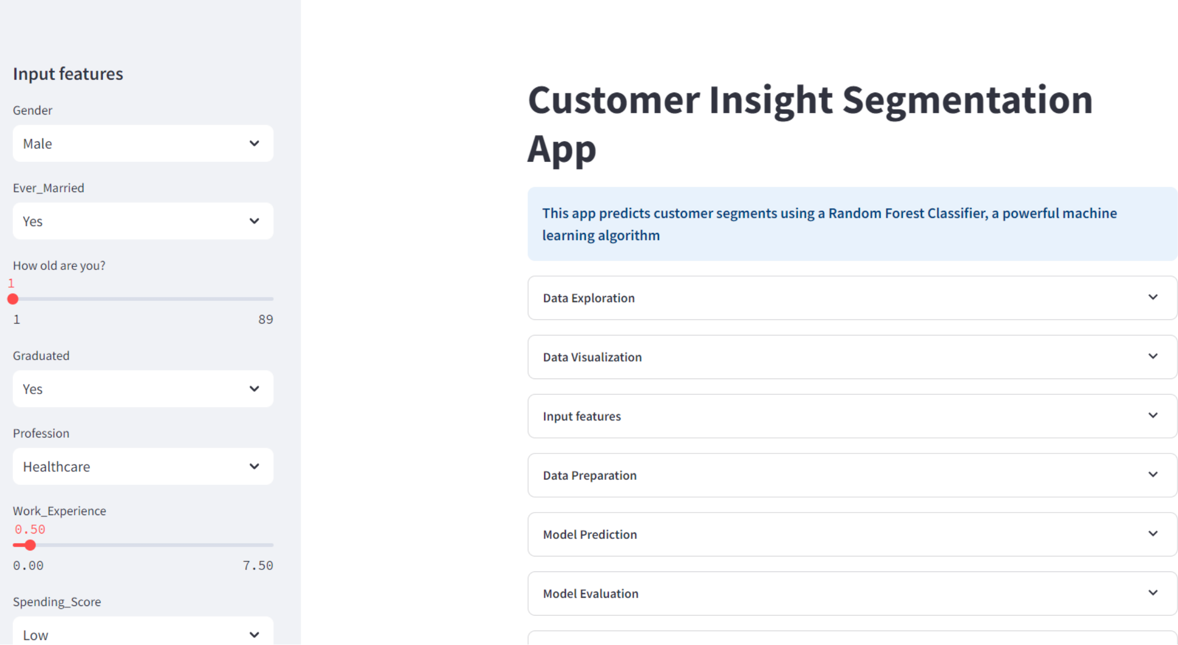
- **Gender: select box [male, female]**
- **Ever married: select box [yes, no]**
- **Age: how old are you [slider 1-89]**
- **Graduated: select box [Yes, NO]**
- **Profession: select box [healthcare, engineer, doctor, lawyer, executive, marketing]**
- **Work_ experience: slider [0-7.5]**
- **Spending score: select box [low, average, high]**
- **Family size: slider [1-9]**
- **Var_1: select box [cat_1 - cat_7]**



```
# Input features
with st.sidebar:


    st.header('Input features')
    Gender=st.selectbox('Gender',('Male','Female'))
    Ever_Married=st.selectbox('Ever_Married',('Yes','No'))
    Age= st.slider('How old are you?',18,89,1)
    Graduated=st.selectbox('Graduated',('Yes','No'))

    Profession=st.selectbox('Profession',('Healthcare ','Entertainment','Engineer','Doctor'
        'Lawyer','Executive','Marketing','Marketing'))
    Work_Experience= st.slider('Work_Experience', 0.0,7.5,0.5)
    Spending_Score= st.selectbox('Spending_Score',('Low','Average','High'))
    Family_Size= st.slider('Family_Size',1,9,1)
    Var_1 = st.selectbox('Var_1',('Cat_1','Cat_2','Cat_3','Cat_4','Cat_5','Cat_6','Cat_7'))
```



Input features

Gender: Male

Ever_Married: Yes

How old are you?: 1 (1 – 89)

Graduated: Yes

Profession: Healthcare

Work_Experience: 0.50 (0.00 – 7.50)

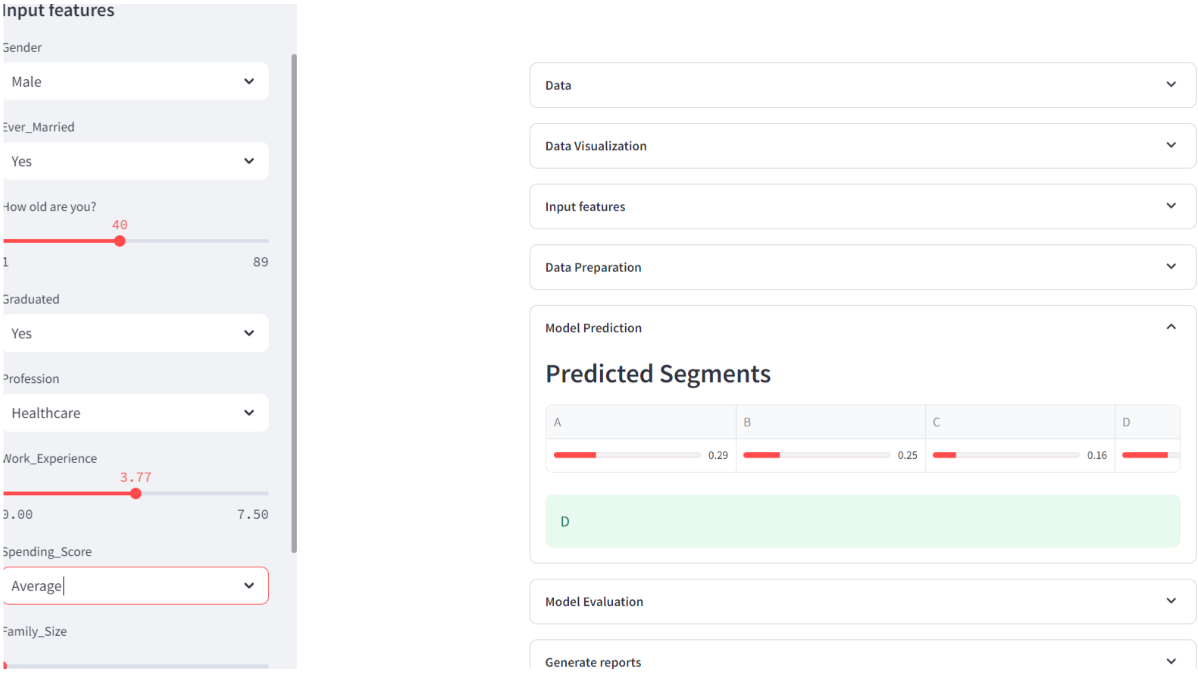
Spending_Score: Low

Customer Insight Segmentation App

This app predicts customer segments using a Random Forest Classifier, a powerful machine learning algorithm

-

7.5 Model prediction: the random forest classifier generates output, the segment of the customer [A, B, C, D] with its probability depending on the input features set by the user



Input features

Gender

Male

Ever_Married

Yes

How old are you?

40

1 89

Graduated

Yes

Profession

Healthcare

Work_Experience

3.77

0.00 7.50

Spending_Score

Average

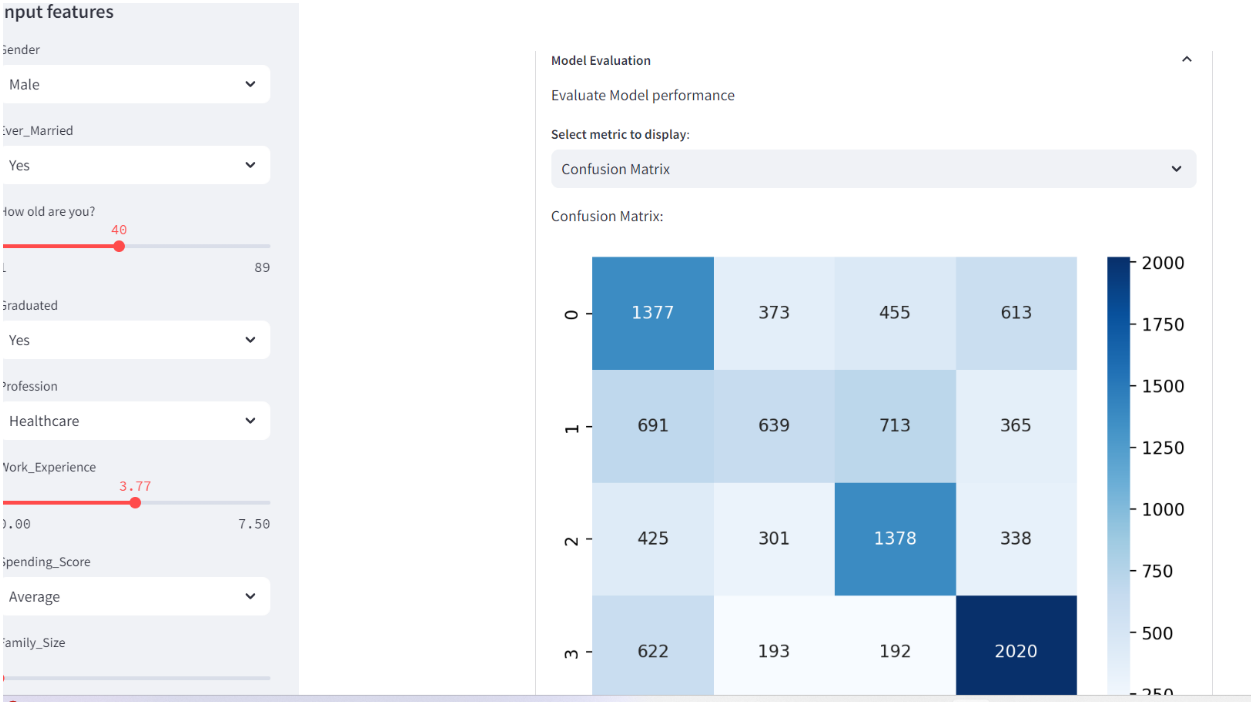
Family_Size

Predicted Segments

A	B	C	D
0.29	0.25	0.16	

D

7.6 Model Evaluation: evaluate model performance using selected metrics: accuracy and confusion matrix from the select box



Input features

Gender

Male

Ever_Married

Yes

How old are you?

40

1 89

Graduated

Yes

Profession

Healthcare

Work_Experience

3.77

0.00 7.50

Spending_Score

Average

Family_Size

Model Evaluation

Evaluate Model performance

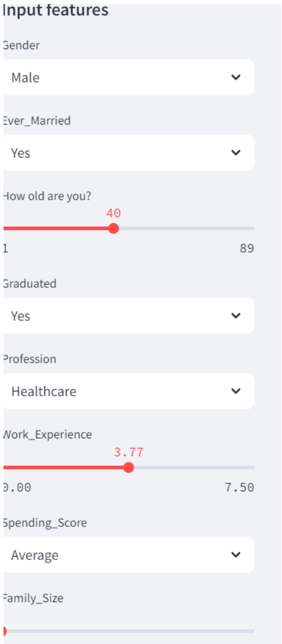
Select metric to display:

Confusion Matrix

Confusion Matrix:

0	1377	373	455	613
1	691	639	713	365
2	425	301	1378	338
3	622	193	192	2020

7.7 **Generate reports :** select input feature from the select box and generate insight using violin charts on how the target variable relates with input features of age, family size and work experience



Input features

Gender

Male

Ever_Married

Yes

How old are you?

40

1 89

Graduated

Yes

Profession

Healthcare

Work_Experience

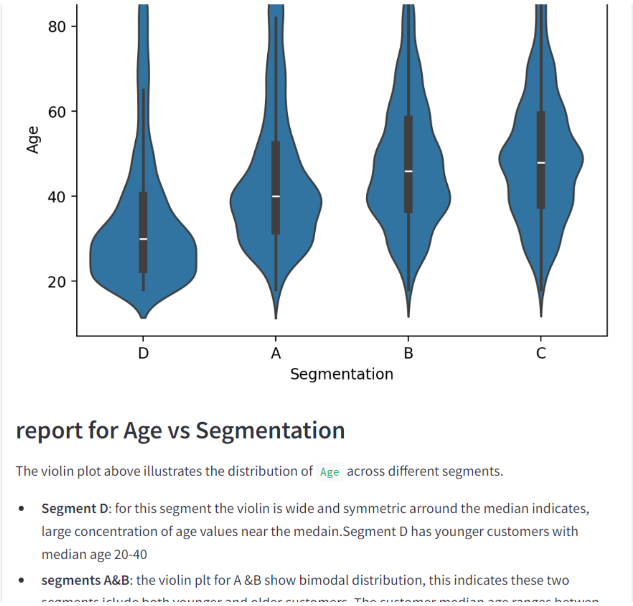
3.77

0.00 7.50

Spending_Score

Average

Family_Size



report for Age vs Segmentation

The violin plot above illustrates the distribution of `Age` across different segments.

- **Segment D**: for this segment the violin is wide and symmetric arround the median indicates, large concentration of age values near the medain.Segment D has younger customers with median age 20-40
- **segments A&B**: the violin plt for A &B show bimodal distribution, this indicates these two segments iclude both younger and older customers. The customer median age ranges betwen

8. **Conclusion**: future enhancements will focus on data presentation, user interface, data visualizations and performance to make the product more intuitive and efficient. Improvements and additions that would made within my data product are as follows:
   - Navigation: creating multi page apps where each page focuses on different content to enhance user experience, and customizable navigation.
   - Data integration: use of streamlit widgets such as ‘st.text_input’, ‘ st.date_input’ to allow users to filter ,select specific dataset from the large dataset
   - Data visualization: adding options for interactive visualizations such as hover effects, zoom to make them more informative
   - Feedback mechanisms: immediate feedback for users for actions like downloading data, plots or filter data
   - Error handling mechanisms: user friendly messages for invalid actions, or processing error

4. Milestone 4: Results Analysis or Testing Components

Components testing

The Component Testing **or** Module Test Cases section focuses on validating the functionality, correctness, and performance of each module in the machine learning pipeline. This is essential to ensure that each part of the system works as expected.

1. **Key Components/ Modules to Test**
    - **Data Input Module**: Allows real-time input of customer data.
    - **Data Preparation Module**: Handles data preprocessing (missing values, scaling, encoding).
    - **Data exploration Module** allows users to preview head of the data, data shape, data information, download data
    - **Model Prediction Module**: Predicts customer segments based on input data.
    - **Model Evaluation Module**: Provides performance metrics like accuracy and confusion matrix.
    - **Data Visualization Module**: Generates visualizations for univariate and bivariate analysis.
    - **Report Generation Module**: Creates and allows users to download analysis reports.
2. **Test Cases for Each Module**
   Each module needs to be tested against both functional and performance criteria.

   A. Data Input Module

   Test case1: valid input data

    - **Scenario**: The user enters valid customer data (e.g., age, income, family size. Gender, ever_married, graduated, etc.)
    - **Expected Output**: The system should accept the input data without errors
    - **Test Execution**: Enter valid data through the input form (e.g. Age=35, family size=2, gender = female, graduated= yes, ever_married = no, profession, executive, work_experience = 2, spending score= average, var_1, cat_2)
    - **Actual Output**: Data is successfully accepted.
    - **Status**: Pass

   Test Case 2: Invalid Input Data

    - **Scenario**: The user enters invalid data (e.g., a string where a predefined selection is expected).

- **Expected Output**: The system should not respond and prompt the user to enter correct data.
- **Test Execution**: Enter invalid data.
- **Actual Output**: system does not respond
- **Status**: Pass

B. Data Preparation Module

Test Case 3: Handling Missing Values

- **Scenario**: The input data contains missing values for certain features.
- **Expected Output**: The system should handle missing values (fill them with mode for categorical features and median for numerical ones).
- **Actual Output**: Missing values are filled or handled as per design.
- **Status**: Pass

Test Case 4: Encoding Categorical Variables

- **Scenario**: The dataset contains categorical variables (such as spending score, segmentation, profession, category, gender, graduated, ever_married) that need to be encoded.
- **Expected Output**: The system should correctly apply ordinal or label encoding.
- **Test Execution**: Provide a dataset with categorical values.
- **Actual Output**: Categorical values are encoded correctly.
- **Status**: Pass

C. Model Prediction Module

Test Case 5: Predicting Customer Segments

- **Scenario**: The system should predict customer segments based on the prepared data.
- **Expected Output**: The system should assign a segment (A, B, C, D) to each customer.
- **Test Execution**: Input customer data and run predictions.
- **Actual Output**: Customers are correctly classified into segments[A,B,C,D].
- **Status**: Pass

Test Case 6: Prediction Speed

- **Scenario**: Test how fast the model generates predictions.
- **Expected Output**: The prediction process should be completed within an acceptable time frame (less than 2 seconds).

- **Test Execution**: Measure the time taken for the model to predict customer segments.
- **Actual Output**: Model prediction time is recorded.
- **Status**: Pass

D. Model Evaluation Module

Test Case 7: Accuracy Metric

- **Scenario**: The system should evaluate the model's accuracy
- **Expected Output**: The system should display the correct accuracy score based on test data.
- **Test Execution**: Run model evaluation on a labeled test set.
- **Actual Output**: Accuracy score is displayed.
- **Status**: Pass

Test Case 8: Confusion Matrix

- **Scenario**: The system should generate a confusion matrix to evaluate model performance.
- **Expected Output**: The confusion matrix should correctly represent true positives, true negatives, false positives, and false negatives.
- **Test Execution**: Run model evaluation and generate the confusion matrix.
- **Actual Output**: Confusion matrix is generated and displayed.
- **Status**: Pass

E. Data Visualization Module

Test Case 9: Univariate Analysis Visualization

- **Scenario**: The system should generate visualizations for single features (histogram of customer age, work experience and family size).
- **Expected Output**: The correct chart should be generated and displayed
- **Actual Output**: Chart is displayed correctly
- **Status**: Pass

Test Case 10: Bivariate Analysis Visualization

- **Scenario**: The system should generate visualizations comparing two features focusing on segmentation (target variable) vs. the categorical variable (e.g. gender vs. segmentation).
- **Expected Output**: The correct bivariate chart(histogram) should be generated.
- **Test Execution**: Run bivariate analysis

- **Actual Output**: Bivariate analysis chart is displayed correctly.
- **Status**: Pass

F. Report Generation Module

Test Case 11: Generating Reports

- **Scenario**: The system should generate a downloadable report summarizing model evaluation, and visualizations.
- **Expected Output**: A downloadable report in png format should be generated.
- **Test Execution**: Generate a report after completing the analysis
- **Actual Output:** Report is generated and downloadable**.**
- **Status**: Pass

Requirements testing

This phase is critical to validate that the system works as intended, and it should map directly to the project's objectives and the functional and non-functional requirements.

1. **Functional Requirements**
   - The system should accept real-time input data for customer behavior.
   - the system allows users to explore the data shape and information and allows users to download data in csv format
   - The system should predict customer segments based on the provided features.
   - The system should provide model evaluation (accuracy, confusion matrix)
   - The system should generate downloadable reports.

2. **Non- functional Requirements**
   - The system should handle requests from computer, mobile simultaneously without failure
   - The interface should be user-friendly and intuitive.
   - The system should comply with data security and privacy

Test Scenario	Test Scenario Description	Expected Outcome	Actual Outcome	Pass/Fail
TS1	Verify that the system can accept	input fields accept valid data	Input was successful; system handled	pass

	customer data inputs		invalid data correctly.	
TS2	Verify the system generate data shape and information and allows users to download data	data shape, info, type and csv data should be displayed properly	Data shape and info displayed and allows users download csv format	Pass
TS3	Verify that the system generates correct univariate and bivariate visualizations.	Univariate/Bivariate charts should display accurate data.	Charts were accurate and consistent with input.	pass
TS4	Ensure that the model generates accurate customer segment predictions	Predictions should match expected customer segments.	Model generated correct customer segments.	Pass
TS5	Verify that the system generates model evaluation matrix	Accuracy and confusion matrix should display accurately	Model generates performance matrix accurately	pass
TS6	Confirm that the user can download the report, and it contains accurate charts.	Download should complete and report should be accurate	Report was downloaded successfully and contained correct charts.	pass
TS7	Verify that the system should handle requests from	The system should work on mobile and desktop properly	System worked on both platforms	pass

	different platforms			
TS8	Verify that system complies data security and privacy	Personal data should anonymize and prevent unauthorized access	System complies with data privacy and security	pass

System testing

This section ensures that the entire system operates as expected, meeting all functional business requirements, business processes, data flows, and other system criteria.

The objective of system testing is to validate that the customer behavior prediction application (Customer Insight Segmentation App) performs as expected across all components, meeting functional business requirements and ensuring seamless integration of processes and data flows.

Test Scenario	Description	Test Steps	Expected Result	Actual Result	Status	Comments
Data Ingestion	Verify that the system can successfully ingest and process customer data from the source.	1. Upload customer data file (CSV). 2. Check data loading status.	The system loads and displays data without errors.	Data successfully loaded and displayed.	Pass	Data uploaded correctly.
Data Preprocessing	Check if data preprocessing (handling missing values, encoding categorical variables) works correctly.	1. input dataset with missing values. 2. Run preprocessing module.	Missing values handled and categorical variables encoded.	Preprocessing completed without issues.	Pass	Data formatted as expected.
Model Prediction	Verify model accuracy in predicting customer segments based on input data.	1. Provide input features. 2. Run the prediction module.	Correct customer segment predicted based on input.	Model correctly predicts the customer segment.	Pass	Model prediction accurate.
Data Visualization	Validate that univariate and bivariate visualizations	1. Select feature for visualization.	Visualizations represent data accurately.	Plots display data correctly.	Pass	Visuals match data expectations.

	are generated accurately.	2. Generate plot.				
Report Generation	Ensure that report generation and downloading functionality works correctly.	Generate a report for the analysis. 2. Download the report.	Report generated and downloaded successfully.	Report downloaded without errors.	Pass	Report includes all relevant details.
Cross-Platform Performance Testing	Verify system performance and responsiveness across different platforms (mobile and desktop).	1. Access the system from a desktop browser. 2. Access the system from a mobile browser. 3. Perform the same operations (e.g., data upload, visualization, prediction) on both platforms.	The system should perform operations smoothly and consistently on both mobile and desktop platforms without lag or UI issues.	System performs well on desktop and mobile platforms with no significant lag or UI discrepancies	Pass	Ensure UI consistency across platforms.
Security Testing	Verify that user authentication and data protection mechanisms are in place.	1. Attempt unauthorized access. 2. Check data encryption.	Unauthorized access denied and data encrypted.	Access denied and data encrypted as expected.	Pass	Security measures effective.
User Acceptance Testing	Ensure that the system meets end-user requirements and expectations.	1. Allow end-users to interact with the system. 2. Collect feedback.	Users should be able to use the system easily and accurately.	Users were able to interact with the system without issues.	Pass	Positive feedback from users.

User Guide

1. **Title page**
   - **Title:** User Guide for Predicting customer behavior using machine learning algorithms: Customer Insight Segmentation App
   - **Date:** 24, September 2024
   - **Author Name:** Abadit Weldeslassie
2. **Table of Contents**

3. **Introduction:**

3.1 Project overview: This project leverages machine learning algorithms to predict customer behavior and segment customers into distinct groups based on their characteristics and actions. It enables businesses to gain deeper insights into customer preferences and make data-driven decisions.

3.2 Core functionality:

- Customer segmentation based on behavioral data
- Predictive analytics for future customer actions
- Data visualization for actionable insights
- Report generation for strategic planning

3.3 Purpose of the System: The purpose of this system is to provide businesses with a comprehensive understanding of their customers, allowing for personalized marketing strategies, improved customer satisfaction, and optimized resource allocation.

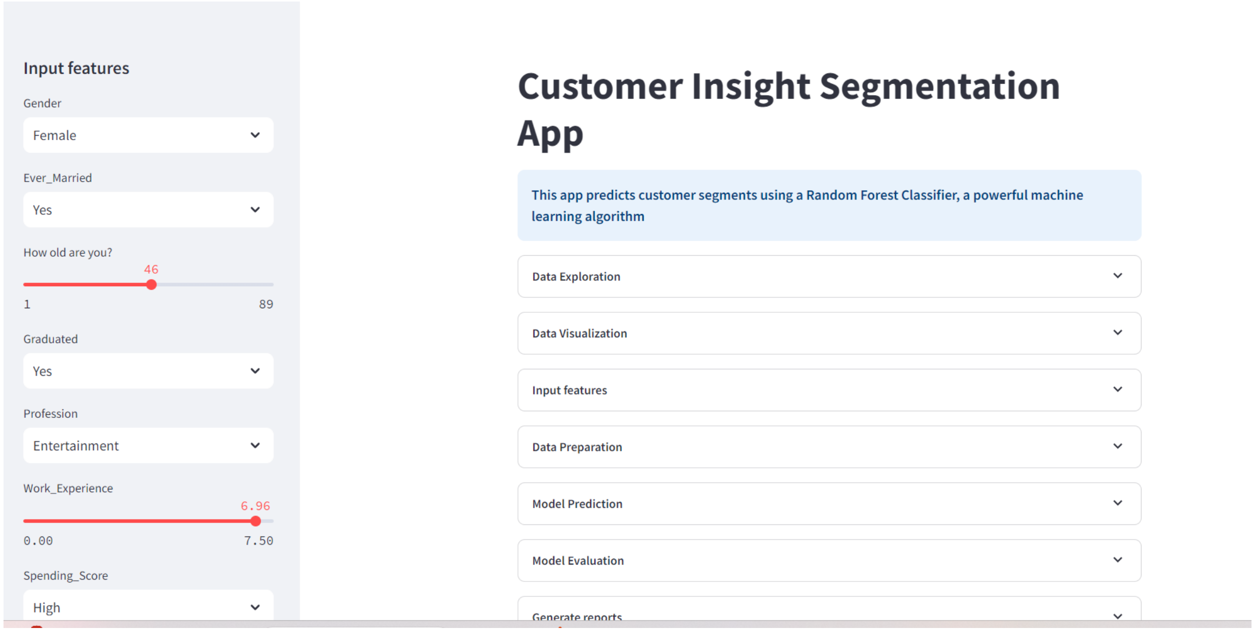
3.4 Target Audience:

- Data Analysts: For conducting detailed customer analysis and generating insights.
- Marketing Teams: For developing targeted campaigns based on customer segments.
- Business Decision Makers: For strategic planning and resource allocation.
- General Users: Anyone interested in understanding customer behavior patterns.

4. **System overview**

4.1 Key features

- Input features
- Data exploration
- Data visualization
- Input features
- Data preparation
- Model prediction
- Model evaluation
- Generate reports



Input features

Gender: Female

Ever_Married: Yes

How old are you? 46 (1 – 89)

Graduated: Yes

Profession: Entertainment

Work_Experience: 6.96 (0.00 – 7.50)

Spending_Score: High

Customer Insight Segmentation App

This app predicts customer segments using a Random Forest Classifier, a powerful machine learning algorithm

5. Navigating the Application

5.1 Input features form (located in the left side of the App) includes the following features

- **Input form fields (located on the left side of the App)**
- **Gender:** Categorical input (Male, Female) in select box
- **Ever married:** Binary input indicating whether the customer is married (Yes/No) in a select box
- **Age:** Numeric input for the customer’s age using slider from 18 to 89 years
- **Graduate: Binary input indicating whether the customer is a graduate** (Yes/No)
- **Profession:** Categorical input for customers profession in a select box(healthcare, engineer, doctor, lawyer, executive, marketing**)**

- **Work experience:** Numeric input for years of work experience using slider 0 to 7.5 years

- **Spending score:** categorical input representing the customer's spending score (low, Average, high) in select box

- **Family size:** Numeric input for the size of the customer's family in slider (1 to 9)

- **Var_1:** anonymized categorical input of the customer in select box (cat_1, cat_2, cat_3, cat_4, cat_5, cat_6, cat_7)

5.2 Data Exploration (Right side of the App):

- **Head of the Raw Data**: Displays the first 10 rows of the dataset to provide a preview.
- **Data Shape**: Shows the dimensions of the dataset (number of rows and columns).
- **Data Info**: Provides a summary of the dataset, including data types and non-null counts.
- **Data Types**: Lists the types of data (e.g., integer, float, categorical) for each column in the dataset.
- **Download CSV Button**: Allows users to download the dataset in CSV format

5.3 Data visualization (right side of the App)

- **Data Summary**: Statistical summary of numeric features (e.g., mean, median, standard deviation).
- **Univariate Analysis:**
- **Histograms**: Visualizations for the distribution of features such as Age, Work Experience, and Family Size. Users can select which feature to visualize from a select box.
- **Value Counts**
- **Value Count Selection**: Allows users to select from features like Gender, Ever Married, Profession, Spending Score, and Graduated. The app generates the value count and displays it along with a bar graph of the selected feature.
- **Bivariate Analysis**:
- **Plotly Graphs**: Interactive Plotly visualizations that show relationships between the target variable (customer segmentation) and input features (Gender, Age, Ever Married, Profession, Spending Score, Graduated). Users select features from a dropdown menu, and the app generates bar graphs based on the selected input.

5.4 Data Preparation (Right Side of the App)

- **Feature engineering**
- **Encoded Input Features**: Displays a table showing the encoded form of input features using encoding methods such as Label Encoding and Ordinal Encoding
- **Encoded Target Variable**: Displays the encoded form of the target variable(segmentation)

5.5 Model Prediction (Right Side of the App)

- **Random Forest Model**: The machine learning model used to predict customer segments based on user input features.
- **Predicted segments:** Displays the probability of each segment (A, B, C, D) in a progress column.
- Shows the predicted customer segment with the highest probability in a single row.

5.6 Model Evaluation

- **Select matrix to display**: Allows users to choose between different evaluation metrics (e.g., accuracy or confusion matrix) for model assessment.
- **Accuracy Metrics**: Provides the accuracy score of the model to assess its performance.
- **Confusion Matrix**: Displays a confusion matrix to evaluate the classification model's performance

5.7 Reports

- **Numerical Feature Insights**
- **Violin Plots**: Visualizes distributions of numerical features (Age, Work Experience, Family Size) against the target variable. Users select features from a dropdown menu.
- **Download and Print Reports**: Users can download or print reports with these visualizations.
- **Categorical Feature Insights**
- **Bar Graphs**: Visualizes distributions of categorical features (Profession, Spending Score, Graduated, Marital Status) against the target variable. Users select features from a dropdown menu.
- **Download and Print Reports**: Users can download or print reports with these visualizations.

5.8 System Interfaces

- **User Interface (UI)**: Layout featuring input fields on the left and functional sections (data exploration, visualization, preparation, prediction, evaluation, and reporting) on the right side.

6. dataset overview

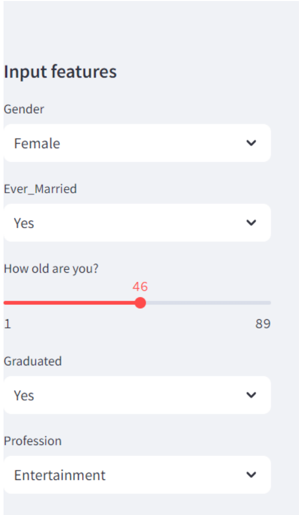
6.1 Dataset description: The dataset is from Kaggle repository, it has two csv files train and test. The train set has 8068 rows and 11 columns, and the test set has 2627 rows and 11 columns

6.2 Source of the data: the dataset is from Kaggle repository, and I did the necessary preprocessing such as

- handling missing values using median for categorical variables and mode for numerical variables
- removing outliers from the profession column using interquartile range (IQR)
- Feature engineering: transforming raw data into features that are suitable for machine learning models and improves the performance of the model using normalization, scaling and encoding

7. step -by step instructions for using the product

7.1 **loading the data:** load the data by clicking the download button



Input features

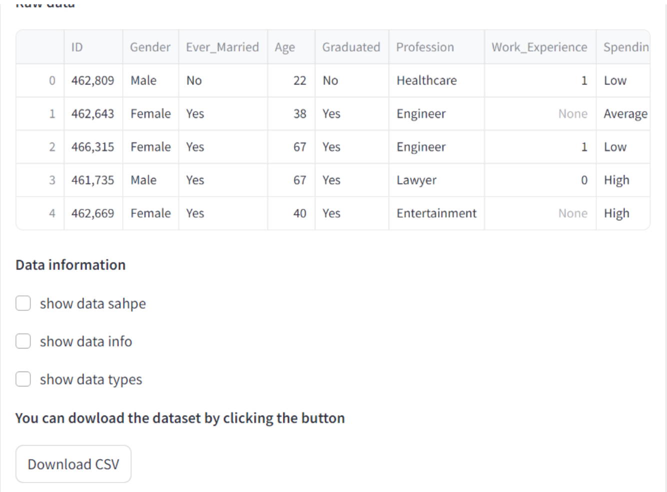
Gender: Female

Ever_Married: Yes

How old are you?: 46 (range 1–89)

Graduated: Yes

Profession: Entertainment



	ID	Gender	Ever_Married	Age	Graduated	Profession	Work_Experience	Spendin
0	462,809	Male	No	22	No	Healthcare	1	Low
1	462,643	Female	Yes	38	Yes	Engineer	None	Average
2	466,315	Female	Yes	67	Yes	Engineer	1	Low
3	461,735	Male	Yes	67	Yes	Lawyer	0	High
4	462,669	Female	Yes	40	Yes	Entertainment	None	High

Data information

□ show data sahpe

□ show data info

□ show data types

You can dowload the dataset by clicking the button

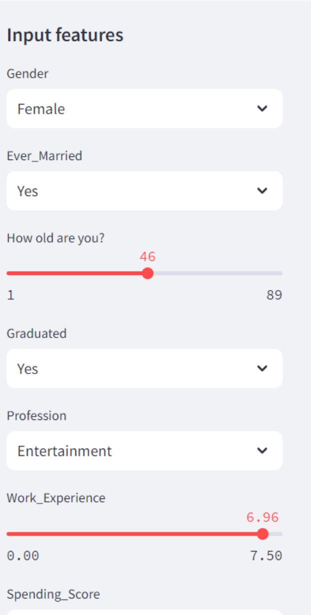
7.2 data exploration: check the shape and information of the data by using show data shape, show data info checkbox

7.3 data visualization: I am visualizing input features and the target variable, it includes

- statistical data summary,

- from the select box select a numerical column for histogram for univariate analysis of age, work experience and family size.

- Form the select box Select a categorical for value count of categorical features
- Bivariate analysis, from the select box select input features such as age, gender etc against the target variable segmentation to understand and explore their relationships



Input features

Gender

Female

Ever_Married

Yes

How old are you?

46

1 89

Graduated

Yes

Profession

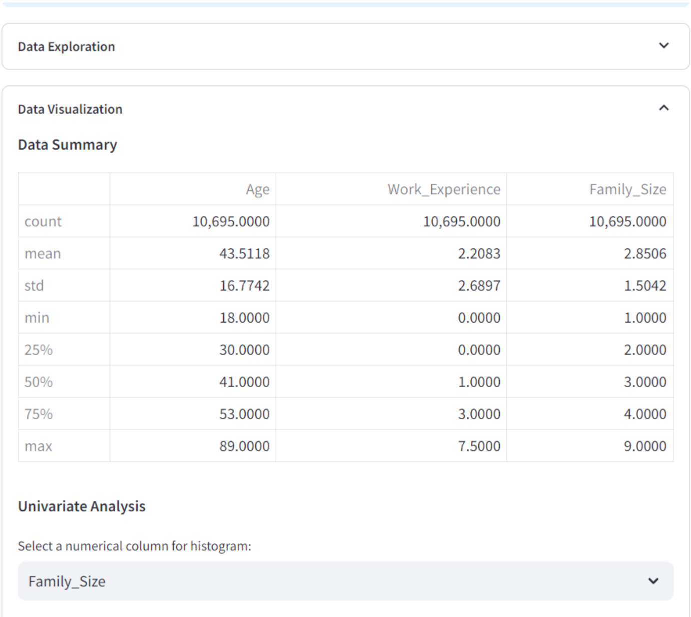
Entertainment

Work_Experience

6.96

0.00 7.50

Spending_Score



Data Exploration

Data Visualization

Data Summary

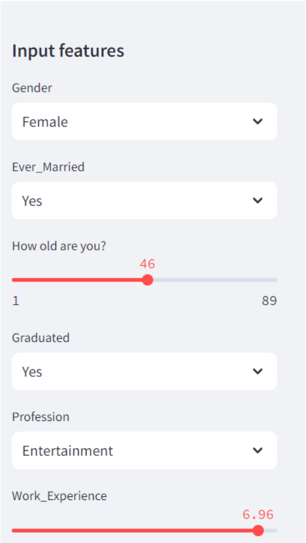
	Age	Work_Experience	Family_Size
count	10,695.0000	10,695.0000	10,695.0000
mean	43.5118	2.2083	2.8506
std	16.7742	2.6897	1.5042
min	18.0000	0.0000	1.0000
25%	30.0000	0.0000	2.0000
50%	41.0000	1.0000	3.0000
75%	53.0000	3.0000	4.0000
max	89.0000	7.5000	9.0000

Univariate Analysis

Select a numerical column for histogram:

Family_Size

7.4 Model prediction: the random forest classifier generates output, the segment of the customer [A, B, C, D] with its probability depending on the input features set by the user



Input features

Gender: Female

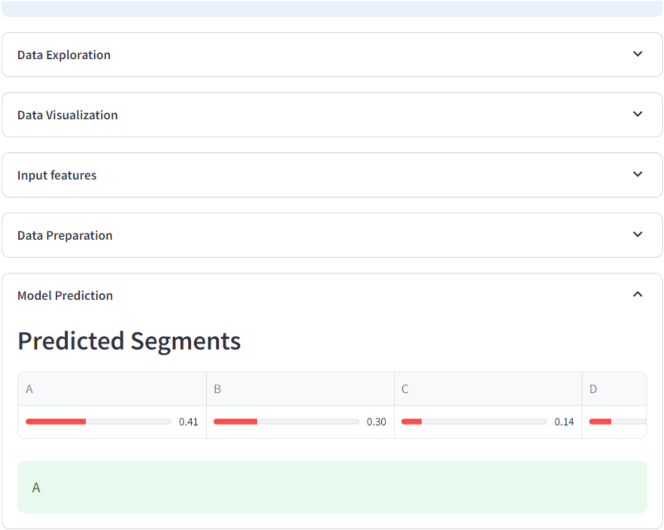
Ever_Married: Yes

How old are you?: 46

Graduated: Yes

Profession: Entertainment

Work_Experience: 6.96

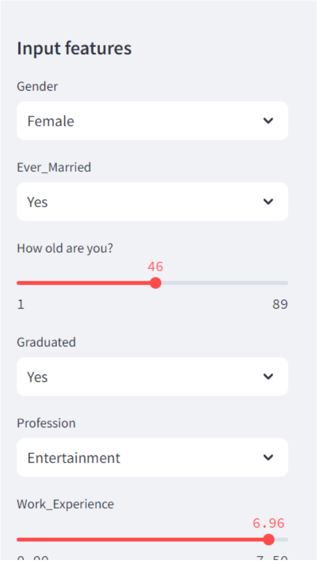


Predicted Segments

A	B	C	D
0.41	0.30	0.14	

A

7.5 Model Evaluation: the random forest classifier generates output, the segment of the customer [A, B, C, D] with its probability depending on the input features set by the user



Input features

Gender: Female

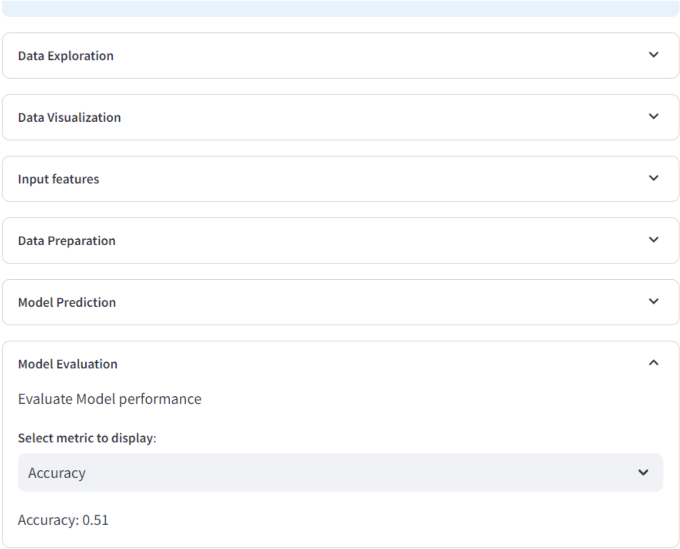
Ever_Married: Yes

How old are you?: 46

Graduated: Yes

Profession: Entertainment

Work_Experience: 6.96



Evaluate Model performance

Select metric to display: Accuracy

Accuracy: 0.51

7.6 **Reports:** select input feature from the select box and generate insight using violin charts on how the target variable relates with input features of age, family size and work experience

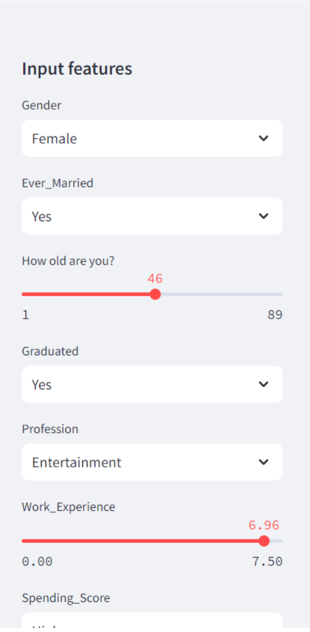
System Administration Guide

1. **Cover page**
    - **Title:** System Administration Guide **for** Predicting customer behavior using machine learning algorithms: Customer Insight Segmentation App
    - **Autho**r: Abadit Weldeslassie
2. **Copy right page**

3. **Table of Contents**

4. **System Overview**

- **Introduction:** predicting customer behavior helps business to better understand their customer needs, preferences and buying pattern, thus helps marketing team to tailor their efforts to reach out customers in the most efficient way for a better customer satisfaction and marketing strategies. The objective of this project is to provide in-depth predictive analysis of customer behavior using machine learning algorithms which plays a vital role in decision making, improve profit rates of business, increase customer satisfaction, and reduce risk by identifying them at the early stage.
- **System Architecture**
    - Input features
    - Data exploration
    - Data visualization
    - Input features
    - Data preparation
    - Model prediction
    - Model evaluation
    - Generate reports



Input features

Gender: Female

Ever_Married: Yes

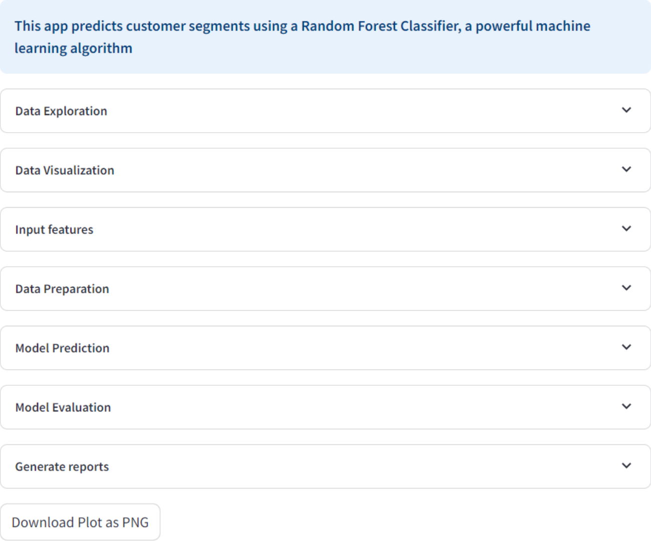
How old are you?: 46 (1 – 89)

Graduated: Yes

Profession: Entertainment

Work_Experience: 6.96 (0.00 – 7.50)

Spending_Score



This app predicts customer segments using a Random Forest Classifier, a powerful machine learning algorithm

- **Technology used**
  - Python
  - Pandas
  - NumPy
  - Matplotlib
  - Plotly
  - Seaborn
  - Scikit-learn
  - Streamlit
  - GitHub
- **System configuration**

  -Hardware requirements

  - CPU: Dual-core processor
  - RAM: 4 GB
  - Storage: 20 GB free disk space
  - Network: Stable internet connection

  - **Software requirements**

    - python 3.8 or higher

    Libraries and packages

    streamlit

    Pandas

    Scikit-learn

    numPy

matplotlib
seaborn

- **System maintenance**
    - Data updates: I will update the dataset regularly to reflect the most recent customer behavior pattern
    - Ensure that data preprocessing and cleaning scripts are up-to-date and functioning as expected.
    - Model maintenance: Schedule regular re-training of the machine learning model with new data to ensure it continues to make accurate predictions.
    - Model performance monitoring: Track the model's performance metrics (accuracy, precision, recall) over time to detect any degradation in performance.
    - Regularly update Python libraries and dependencies to their latest versions to maintain compatibility and security.
    - Regularly review and update the Streamlit app's user interface based on user feedback to improve the user experience.

- **Security related process**
    - **Data Anonymization:** the creator of the dataset already implemented strong anonymization techniques to protect individuals' privacy. This includes removing or masking personally identifiable information (PII) to ensure that the data cannot be traced back to specific individuals.
    - **Transparency and Consent:** Although my project used publicly available datasets, I considered the importance of transparency and informed consent in data collection.
    - **Respecting Ownership:** I ensured that all third-party content, including datasets, algorithms, and code, was used in accordance with their respective licenses.

References

Practical data science. A quick guide to customer segmentation for data scientists.

Practical data science. https://practicaldatascience.co.uk/data-science/a-quick-guide-to-customer-segmentation

Sabbeh, S.F. (2018). Machine-Learning Techniques for Customer Retention: A Comparative Study. *International Journal of Advanced Computer Science and Applications, 9*.

Asniar and K. Surendro.(2019).Predictive Analytics for Predicting Customer Behavior. International Conference of Artificial Intelligence and Information Technology (ICAIIT). pp. 230-233.

DOI: 10.1109/ICAIIT.2019.8834571

T. Kansal, S. Bahuguna, V. Singh and T. Choudhury. (2018). Customer Segmentation using K-means Clustering. *International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*. pp. 135-139, doi: 10.1109/CTEMS.2018.8769171.

https://www.kaggle.com/datasets/kaushiksuresh147/customer-segmentation/data

vidya, Analytics. (January 8, 2021). OSEMN is Awesome. Medium.
https://medium.com/analytics-vidhya/osemn-is-awesome-3c9e42c3067d

Sharma, Udit (May 29, 2024). OSEMN Framework for Data Science. LinkedIn
https://www.linkedin.com/pulse/osemn-framework-data-science-pronounced-awesome-udit-sharma-26blc/

Elbert, Christof. Data Science: Technologies for Better Software**.** Software Technology.**https://ieeexplore-ieee-org.lopes.idm.oclc.org/stamp/stamp.jsp?tp=&arnumber=8880036**

(July 03, 2024). Evaluation metrics in machine learning. Geeksforgeeks.
https://www.geeksforgeeks.org/metrics-for-machine-learning-model/

Kumar, Dhairya. (December 25, 2018). Introduction to data preprocessing in machine learning. Medium. https://towardsdatascience.com/introduction-to-data-preprocessing-in-machine-learning-a9fa83a5dc9d