Predicting Customer Behavior Using Machine Learning

Customer Insight Segmentation App

- **Abadit Weldeslassie**

- **October 7, 2024**

Introduction

- Customer behavior analysis is crucial for businesses to better understand their customers' needs, preferences, and buying patterns. This understanding enables marketing teams to tailor their strategies more effectively, ensuring efficient outreach, increased customer satisfaction, and enhanced marketing strategies.

- Machine learning and predictive analysis are essential tools for uncovering insights from vast amounts of data, helping businesses identify customer needs and behavior trends. With the growing availability of customer data from sources like online transactions, social media, and customer service interactions, traditional analysis methods are becoming insufficient. As businesses struggle to extract meaningful insights, they face missed opportunities and inefficiencies in targeting and customer engagement efforts.

Scope

Objective: 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.

Problem Statement : Understanding customer behavior is crucial for:

- **Personalized marketing**: leading to a more satisfying customer experience
- **Efficient resource allocation**: which maximize revenue Increase
- **customer retention**: increasing loyalty and reducing churn

Design and Foundation

Data Description: I am using a dataset from Kaggle repository . 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.

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.

Data overview

- **Head of the clean data**

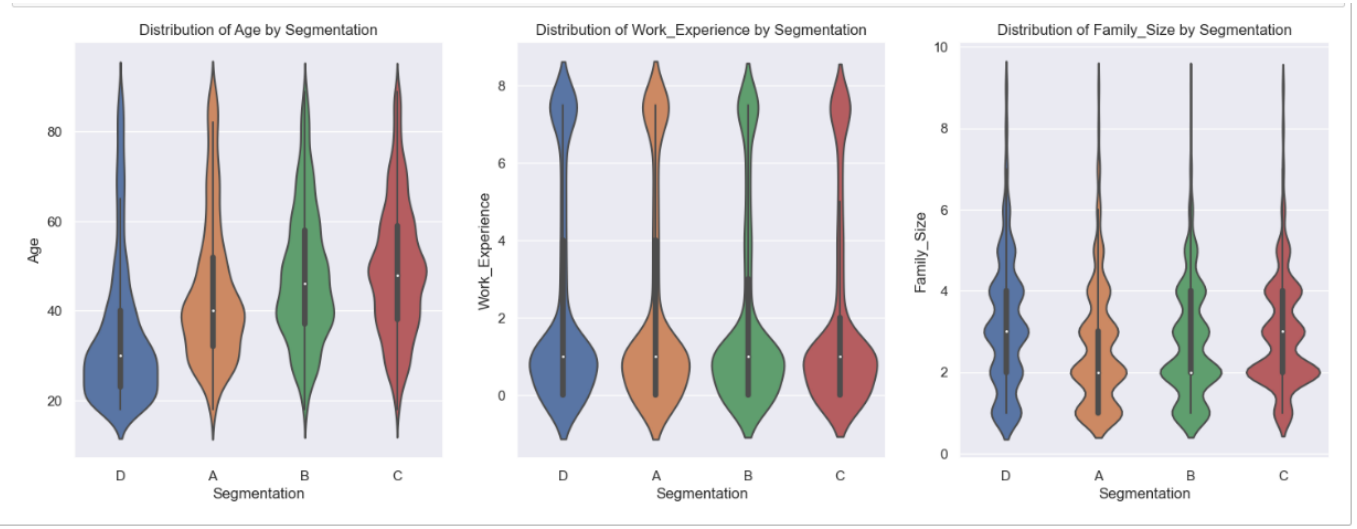
	Gender	Ever_Married	Age	Graduated	Profession	Work_Experience	Spending_Score	Family_Size	Var_1	Segmentation
0	Male	No	22	No	Healthcare	1	Low	4	Cat_4	D
1	Female	Yes	38	Yes	Engineer	1	Average	3	Cat_4	A
2	Female	Yes	67	Yes	Engineer	1	Low	1	Cat_6	B
3	Male	Yes	67	Yes	Lawyer	0	High	2	Cat_6	B
4	Female	Yes	40	Yes	Entertainment	1	High	6	Cat_6	A

- **Summary statistics**

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

Exploratory Data Analysis (Bivariate Analysis)

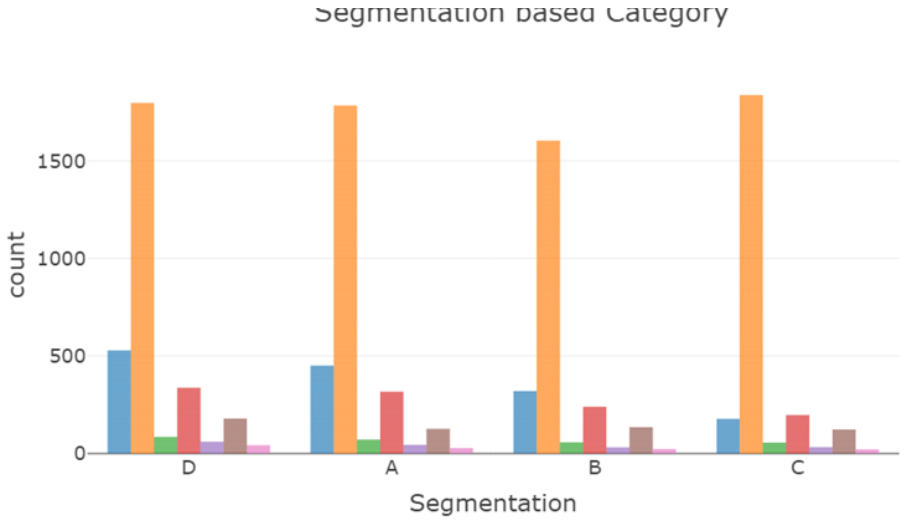


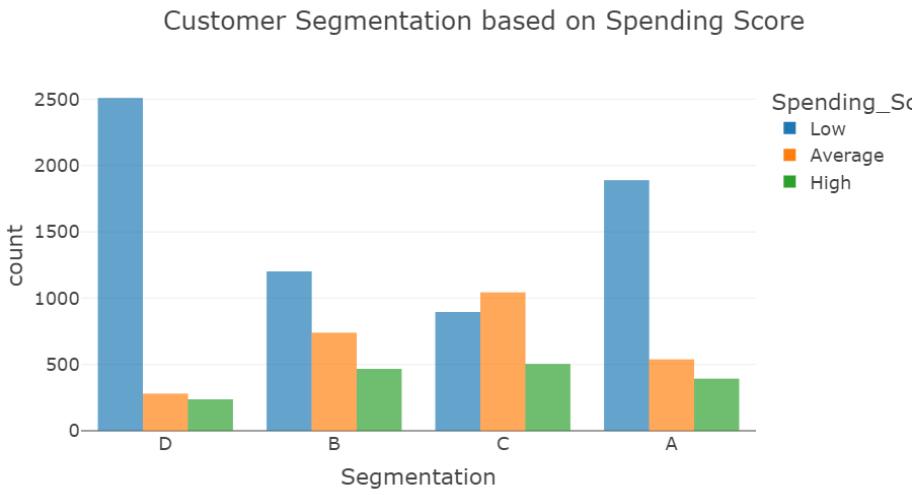
The following observations can be derived from the plots

- **Age Range for *Segment A, B, C, D*:** The distribution is spread from around 20 to 80+ years
- **Median Age**: The median age for the fours seems to be closer to 40
- Segment **D**: Density: There is a slightly higher density around the ages of 30-40 and another around 50-60, indicating two subgroups within this segment.
- **Segment A, Density**: The distribution is slightly more concentrated around the ages of 30-40, meaning more individuals in this segment are younger or middle-aged.
- **Density for Segment C and B** appears to have a relatively even distribution of ages, with slight concentrations in the middle-age range (30-50).

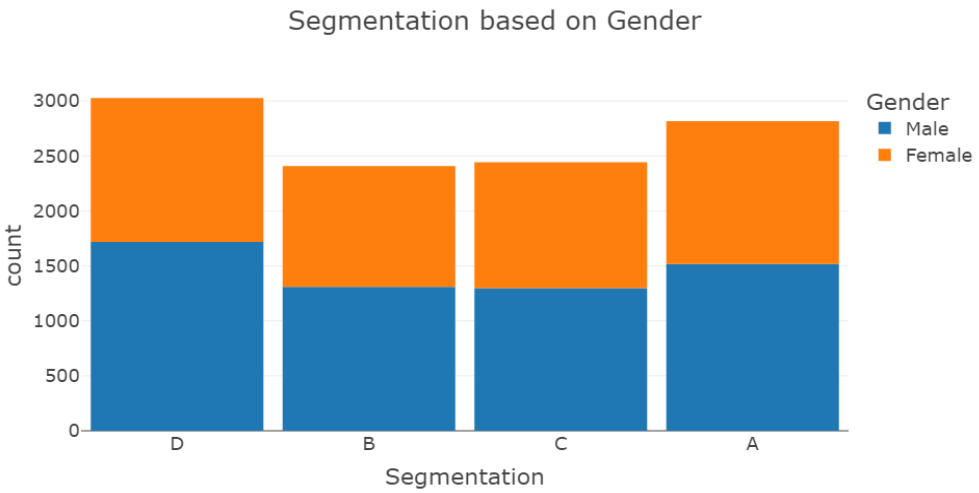
Customers in Segments D and A tend to have low spending scores, suggesting they may need different strategies to increase their spending.

Segments B and C appear more diverse in spending behavior, indicating a potential to target these groups for upselling or personalized offers

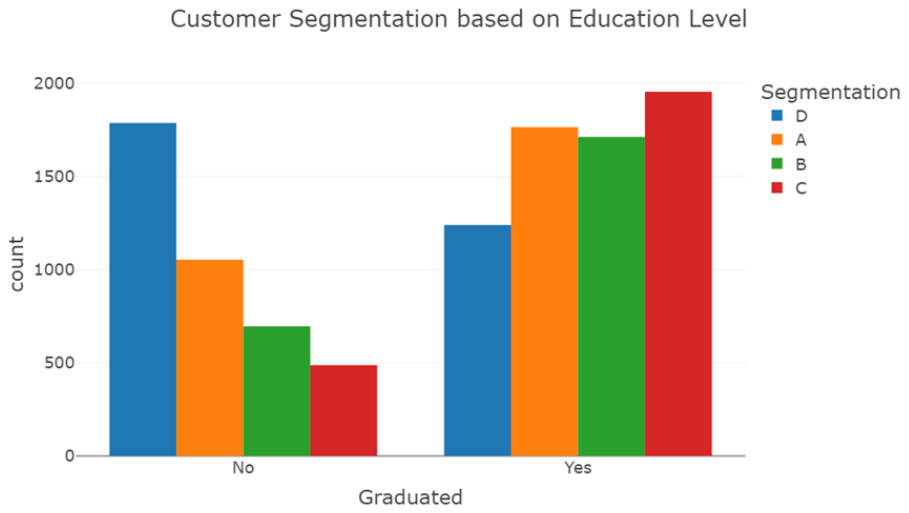




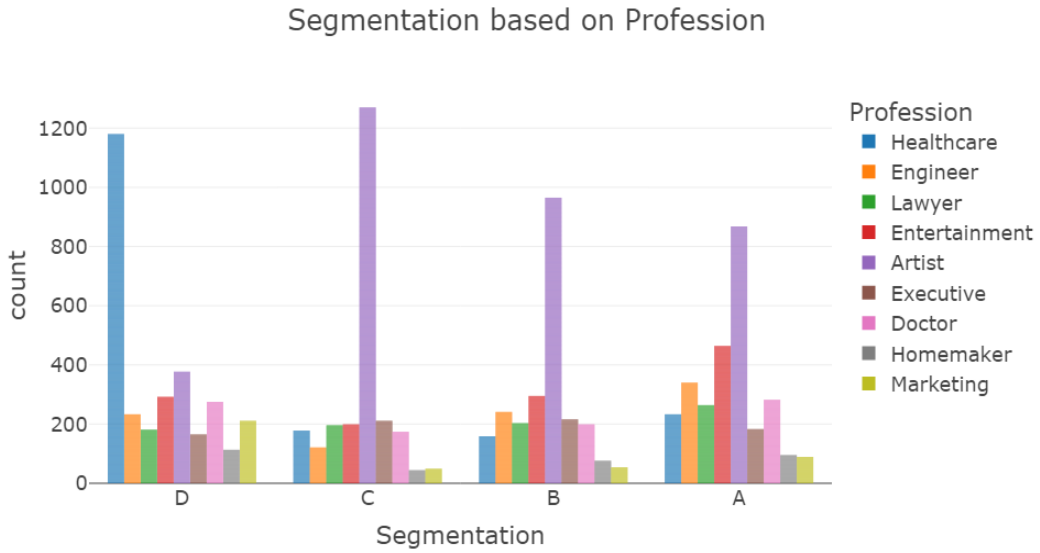
Segment D is more dominant among non-graduates, suggesting that customers without a higher level of education may have different behaviors or preferences that align them with this segment.
Segments A, B, and C are more common among graduates, indicating that education level might positively influence customer behavior, leading them to higher-value segments

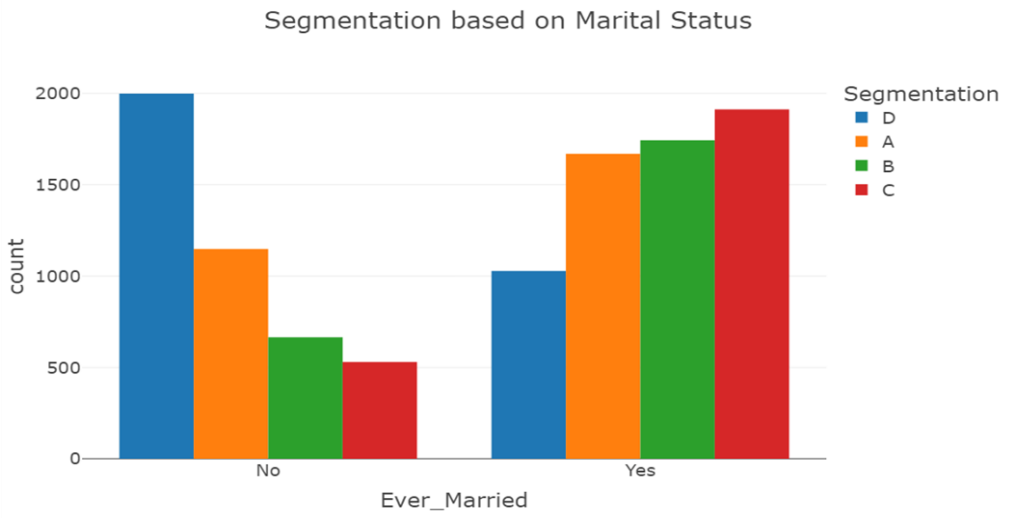


Segment D has the highest count of both male and female customers
The gender distribution across Segments B, C, and A is balanced, with a similar proportion of male and female customers in each.
The consistent ratio of males to females in most segments suggests that gender might not be a significant differentiator in customer behavior for these segments



Segment C (Lowest Unmarried, Highest Married):
Family-Oriented Campaigns: Focus on family-friendly products or services. Highlight benefits that cater to family life, such as discounts for family outings or family packs
Segment D (Highest Unmarried, Lowest Married)
Single Lifestyle Promotion: Tailor marketing messages to celebrate the single lifestyle. Promote products or experiences that enhance social activities and personal growth.
Segment A, B, C (Mostly Artists, Least Marketing)
Visual Marketing: Use visually appealing content in campaigns that resonate with the artistic sensibilities of these segments. Consider using platforms like Instagram and P **Tailored Offers**: Create offers specifically for artists, such as discounts on art supplies or tools. Highlight how your products can enhance their creativity
Segment D (Dominated by Healthcare, Least Homemaker)
Health-Focused Products: Tailor marketing messages around wellness, self-care, and preventive health. Highlight products that promote a healthy lifestyle.
Partnerships with Healthcare Providers: Collaborate with clinics or wellness centers to offer exclusive promotions, further establishing your brand in the healthcare community





Design and Foundation

Data preprocessing :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: Feature engineering is the process of transforming raw data into features that are suitable for machine learning models. Its main aim is to improve model accuracy by providing more meaningful and relevant information. This includes:

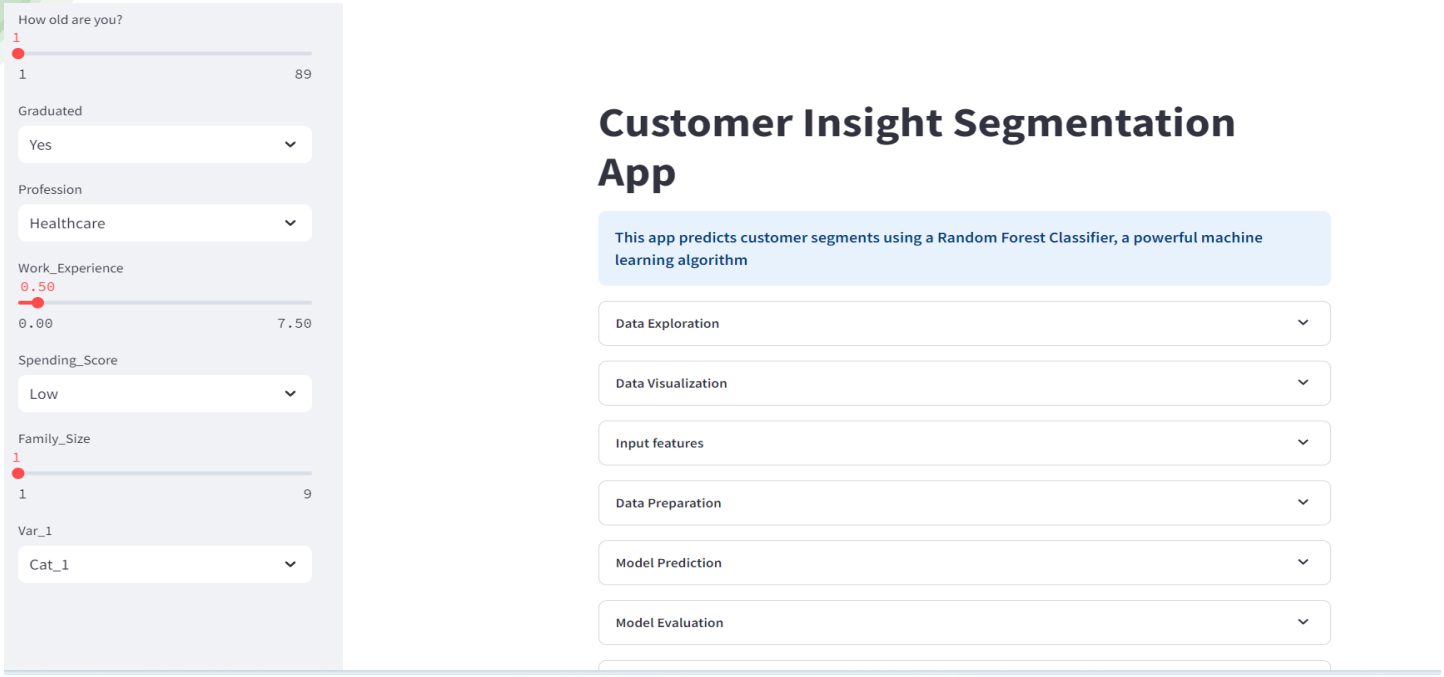
- **Encoding** :
- Convert categorical data (segmentation, profession, category) into numerical format using label encoding
- Convert categorical columns (spending score) into numerical format using ordinal encoding
- Convert categorical columns (gender, graduated, ever married) into numerical format using predefined mapping
- **Feature selection :** I use Mutual information (MI) metric to assess the importance or relevance of features for the model

Machine learning models

Model	Description	Model Accuracies:	Selection Reason
Logistic Regression	Used for binary classification to predict customer segment	44.55%	Chose **Random Forest** for its robustness and ability to handle overfitting effectively, despite being ranked fourth .
Decision Tree	Provides a clear visualization of decision-making criteria.	37.4%	
K-Nearest Neighbors	Identifies segments based on similarity to other customers.	46%	
Support Vector Machine (SVM)	Effective for high-dimensional data, separating classes with hyperplanes.	47.64%	
Random Forest	Combines multiple decision trees for improved accuracy and robustness.	41.75%	

Models are compared on accuracy , precision, recall ,f1-score

Customer Insight Segmentation App Interface



Customer Insight Segmentation App Overview

- **Input Features (Left Side):**Users can manipulate:
- Gender, Age, Graduated, Ever Married
- Profession, Work Experience, Spending Score
- Family Size, Var_1
- **Key Activities (Right Side**):

1. **Data Exploration**:

- View data shape, type, and information.
- Download dataset as CSV

2. Data Visualization:

- Summary statistics and univariate analysis (histograms).
- Value counts in table and bar graph formats.
- Bivariate analysis charts for target vs. input features.

3. Data Preparation:

- View encoded input features

4. Model Prediction:

- Random Forest model predicts customer segments (A, B, C, D).
- Displays probabilities and highest probability segment.

5. Model Evaluation:

- Choose evaluation metrics (e.g., accuracy, confusion matrix).
- Display model performance.

6. Reports:

- Download univariate and bivariate plots with reports.

Business Impact of Customer Insight Segmentation App

- **Purpose of the System**:

- Comprehensive Customer Understanding:
- Enables personalized marketing strategies.
- Enhances customer satisfaction.
- Optimizes resource allocation

- **Target Audience**:
- Data Analysts:
- conduct detailed customer analysis and generate insights
- Marketing Teams:
- Develop targeted campaigns based on customer segments
- Business Decision Makers:
- Facilitate strategic planning and resource allocation
- General Users:
- Explore customer behavior patterns.

Conclusion

- **Effective Segmentation**:
- Machine learning models accurately predict customer segments, enabling tailored marketing strategies
- **User-Friendly Design**:
- Intuitive app empowers users to gain insights without technical expertise
- **Limitations:**
- Data Quality: Insights depend on data completeness and accuracy
- **Future Improvements**:
- Hyperparameter Tuning: explore different settings for the Random Forest model to find optimal configurations.
- Investigate any underrepresented segments; consider oversampling or under sampling techniques.

Acknowledgments

- Thank you to my GCU instructors for their guidance, support, and valuable insights throughout this project.
- A special thanks to my classmates for their collaboration, feedback, and encouragement.
- Heartfelt gratitude to my family for their unwavering support and motivation during this journey.

References

- 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
- 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