Customer Segmentation Analysis and Customer Churn Prediction

Yiming Liu	Jinhua Yan	Wenxin Dong	Vikki Wong	Hadya Adnan
400419748	400428258	400429070	001318521	400359567
Software Engineering	Software Engineering	Software Engineering	Software Engineering	Software Engineering
McMaster University	McMaster University	McMaster University	McMaster University	McMaster University
London, Ontario	Toronto, Ontario	Toronto, Ontario	Toronto, Ontario	Toronto, Ontario

Abstract

This report delves into customer segmentation analysis and customer churn prediction using a dataset containing customer demographics, subscription details, and behavior. The objective is to segment customers into distinct groups based on their characteristics. The analysis utilizes machine learning techniques, including K-Means clustering and decision tree classification. The findings provide valuable insights for businesses to improve customer retention and tailor marketing strategies.

I. INTRODUCTION

Customer Segmentation Analysis and Customer Churn Prediction are crucial tasks in customer analytics and business intelligence. These processes help businesses understand their customers better, identify distinct customer groups, and predict which customers will likely churn, i.e., stop using their services.

This report shows the steps taken to perform Customer Segmentation Analysis and Customer Churn Prediction based on a dataset from Kaggle containing customer information. The analysis uses Python programming language and popular data science libraries such as NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn, and Imbalanced-learn.

The heart of this analysis lies in Customer Segmentation, where the Elbow Method helps determine the optimal number of customer clusters. Using K-means clustering, we group customers into distinct segments based on shared characteristics. Additionally, we present a cluster analysis heatmap to uncover underlying patterns in the data.

We employ machine learning models to predict customer churn, including Logistic Regression, Decision Tree, Random Forest, and Neural Network (MLP). We evaluate their performance to identify the most accurate and interpretable model. Ultimately, the Decision Tree model emerges as the chosen model, providing insights into feature importance and creating a readable representation for easy understanding.

Insights from Customer Segmentation Analysis and Customer Churn Prediction empower businesses to take proactive measures in customer engagement and retention. By leveraging data-driven strategies, companies can enhance customer experience, reduce churn rates, and foster long-lasting customer loyalty, bolstering overall business growth and success.

	CustomerID	Age	Gender	Tenure	Usage Frequency	Support Calls	Payment Delay	Subscription Type	Contract Length	Total Spend	Last Interaction	Churn
0	2.0	30.0	Female	39.0	14.0	5.0	18.0	Standard	Annual	932.00	17.0	1.0
1	3.0	65.0	Female	49.0	1.0	10.0	8.0	Basic	Monthly	557.00	6.0	1.0
2	4.0	55.0	Female	14.0	4.0	6.0	18.0	Basic	Quarterly	185.00	3.0	1.0
3	5.0	58.0	Male	38.0	21.0	7.0	7.0	Standard	Monthly	396.00	29.0	1.0
4	6.0	23.0	Male	32.0	20.0	5.0	8.0	Basic	Monthly	617.00	20.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...	...
440828	449995.0	42.0	Male	54.0	15.0	1.0	3.0	Premium	Annual	716.38	8.0	0.0
440829	449996.0	25.0	Female	8.0	13.0	1.0	20.0	Premium	Annual	745.38	2.0	0.0
440830	449997.0	26.0	Male	35.0	27.0	1.0	5.0	Standard	Quarterly	977.31	9.0	0.0
440831	449998.0	28.0	Male	55.0	14.0	2.0	0.0	Standard	Quarterly	602.55	2.0	0.0
440832	449999.0	31.0	Male	48.0	20.0	1.0	14.0	Premium	Quarterly	567.77	21.0	0.0

440833 rows × 12 columns

Fig. 1 Overview of the dataset

II. PREPROCESS

A. The Data Set

The data set for this project is downloaded from Kaggle. The dataset comprises a diverse range of customer-related features, including both numerical and categorical data. The dataset's overview is shown below in Fig. 1.

B. Data Cleaning

Data cleaning is crucial in data analysis, ensuring the dataset is accurate, consistent, and free from errors and missing values. This section will outline the essential steps taken during the data cleaning process for our Customer Segmentation Analysis and Customer Churn Prediction.

Missing values in the dataset can lead to biased analysis and inaccurate predictions. We address this issue by identifying and handling missing data.

The dataset is clean for analysis after using Panda's method dropna() to remove any rows with missing values. The dataset info is shown in Fig 2.

The next step is converting non-numeric data into numerical data. Some columns in the dataset may contain non-numeric data, such as categorical variables. We convert these non-numeric values into numeric form to make the data suitable for analysis. This conversion is achieved by mapping the non-numeric categories to numeric representations using dictionaries. Figures 3 and 4 show before and after converting.

```
Index: 440832 entries, 0 to 440832
Data columns (total 12 columns):
 #   Column             Non-Null Count   Dtype
---  ------             --------------   -----
 0   CustomerID         440832 non-null  float64
 1   Age                440832 non-null  float64
 2   Gender             440832 non-null  int64
 3   Tenure             440832 non-null  float64
 4   Usage Frequency    440832 non-null  float64
 5   Support Calls      440832 non-null  float64
 6   Payment Delay      440832 non-null  float64
 7   Subscription Type  440832 non-null  int64
 8   Contract Length    440832 non-null  int64
 9   Total Spend        440832 non-null  float64
 10  Last Interaction   440832 non-null  float64
 11  Churn              440832 non-null  float64
dtypes: float64(9), int64(3)
memory usage: 43.7 MB
```

Fig. 2 After removing missing values.

Mapping Rules:

- Gender: 0 represents Female, and 1 represents Male.
- Subscription: 0 represents Standard, 1 represents Basic, and 2 represents Premium.
- Contract Length: 0 represents Monthly, 1 represents Quarterly, and 2 represents Annual.

```
CustomerID [2.00000e+00 3.00000e+00 4.00000e+00 ... 4.49997e+05 4.49998e+05
 4.49999e+05]
Age [30. 65. 55. 58. 23. 51. 39. 64. 29. 52. 22. 48. 24. 49. 19. 47. 42. 57.
 27. 59. 21. 60. 35. 18. 56. 20. 63. 25. 28. 32. 38. 37. 31. 53. 41. 33.
 26. 36. 44. 34. 61. 40. 45. 46. 54. 43. 50. 62.]
Gender ['Female' 'Male']
Tenure [39. 49. 14. 38. 32. 33. 37. 12.  3. 18. 21. 41. 35.  4. 56. 44. 15. 55.
 43. 52. 26.  2. 29. 59. 40. 51. 53. 24. 30.  6. 28. 17. 60.  7. 34. 10.
  5. 45. 54. 58. 25. 13. 47. 31. 22. 19. 23.  1.  8. 46. 16. 50. 48. 11.
 42. 27.  9. 20. 57. 36.]
Usage Frequency [14.  1.  4. 21. 20. 25. 12.  8.  5.  9.  6. 17. 23. 13. 16. 27.  2. 28.
 29. 15. 24.  3. 22. 26. 30.  7. 11. 18. 19. 10.]
Support Calls [ 5. 10.  6.  7.  9.  3.  4.  2.  0.  1.  8.]
Payment Delay [18.  8.  7. 26. 16. 15.  4. 11. 30. 25. 13. 22.  5. 14.  3. 10. 28.  2.
  6. 27. 12. 29. 17. 24.  9. 23. 21.  1.  0. 20. 19.]
Subscription Type ['Standard' 'Basic' 'Premium']
Contract Length ['Annual' 'Monthly' 'Quarterly']
Total Spend [932.   557.   185.   ... 829.59 804.3  959.47]
Last Interaction [17.  6.  3. 29. 20.  8. 24. 30. 13. 18. 19. 23.  4. 16. 10. 21. 22.  2.
 15. 28. 26.  7.  1.  9.  5. 14. 11. 12. 27. 25.]
Churn [1. 0.]
```

Fig. 4 Before converting non-numeric data to numerical data.

```
CustomerID [2.00000e+00 3.00000e+00 4.00000e+00 ... 4.49997e+05 4.49998e+05
 4.49999e+05]
Age [18. 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29. 30. 31. 32. 33. 34. 35.
 36. 37. 38. 39. 40. 41. 42. 43. 44. 45. 46. 47. 48. 49. 50. 51. 52. 53.
 54. 55. 56. 57. 58. 59. 60. 61. 62. 63. 64. 65.]
Gender [0 1]
Tenure [ 1.  2.  3.  4.  5.  6.  7.  8.  9. 10. 11. 12. 13. 14. 15. 16. 17. 18.
 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29. 30. 31. 32. 33. 34. 35. 36.
 37. 38. 39. 40. 41. 42. 43. 44. 45. 46. 47. 48. 49. 50. 51. 52. 53. 54.
 55. 56. 57. 58. 59. 60.]
Usage Frequency [ 1.  2.  3.  4.  5.  6.  7.  8.  9. 10. 11. 12. 13. 14. 15. 16. 17. 18.
 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29. 30.]
Support Calls [ 0.  1.  2.  3.  4.  5.  6.  7.  8.  9. 10.]
Payment Delay [ 0.  1.  2.  3.  4.  5.  6.  7.  8.  9. 10. 11. 12. 13. 14. 15. 16. 17.
 18. 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29. 30.]
Subscription Type [0 1 2]
Contract Length [0 1 2]
Total Spend [ 100.    100.02  100.06 ...  999.98  999.99 1000.  ]
Last Interaction [ 1.  2.  3.  4.  5.  6.  7.  8.  9. 10. 11. 12. 13. 14. 15. 16. 17. 18.
 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29. 30.]
Churn [0. 1.]
```

Fig. 3 After converting non-numeric data to numerical data.

C. Correlation Analysis

Correlation Analysis is a statistical technique used to quantify and understand the relationship between two or more numerical variables in a dataset. It helps us identify how changes in one variable are associated with changes in another variable.

- A positive correlation ($r > 0$) indicates that as one variable increases, the other also increases, and vice versa.
- A negative correlation ($r < 0$) implies that as one variable increases, the other decreases, and vice versa.
- A correlation close to zero ($r \approx 0$) suggests that there is no significant linear relationship between the variables.

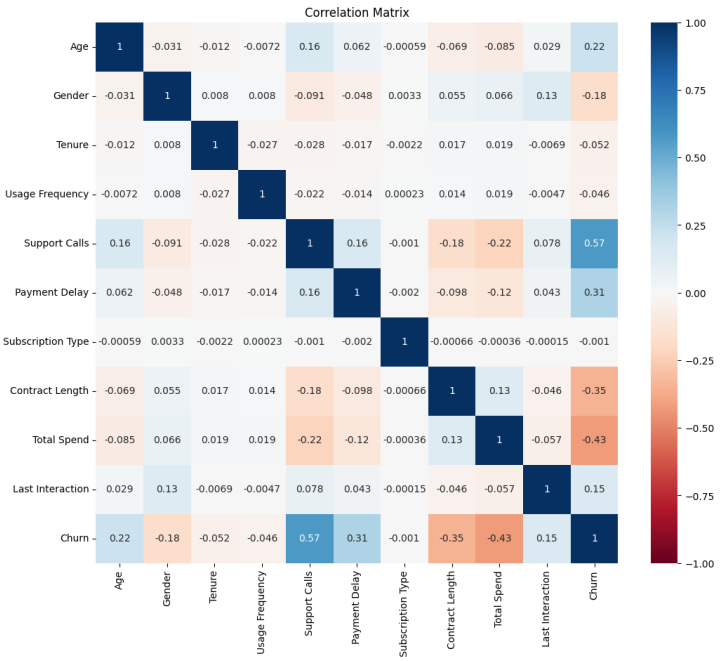

Fig. 5 Correlation Matrix

D. Column Features

Select the all-column features except the Customer ID and use describe() function to get the key statistics and insights about each feature in the dataset. From Fig 6, we can get some interpretations of features.

	count	mean	std	min	25%	50%	75%	max
Age	440832.0	39.373153	12.442369	18.0	29.0	39.0	48.0	65.0
Gender	440832.0	0.567681	0.495399	0.0	0.0	1.0	1.0	1.0
Tenure	440832.0	31.256336	17.255727	1.0	16.0	32.0	46.0	60.0
Usage Frequency	440832.0	15.807494	8.586242	1.0	9.0	16.0	23.0	30.0
Support Calls	440832.0	3.604437	3.070218	0.0	1.0	3.0	6.0	10.0
Payment Delay	440832.0	12.965722	8.258063	0.0	6.0	12.0	19.0	30.0
Subscription Type	440832.0	0.998979	0.821921	0.0	0.0	1.0	2.0	2.0
Contract Length	440832.0	1.204373	0.746851	0.0	1.0	1.0	2.0	2.0
Total Spend	440832.0	631.616223	240.803001	100.0	480.0	661.0	830.0	1000.0
Last Interaction	440832.0	14.480868	8.596208	1.0	7.0	14.0	22.0	30.0
Churn	440832.0	0.567107	0.495477	0.0	0.0	1.0	1.0	1.0

Fig. 6 Describe on column features.

- The average age of customers is around 39 years with a standard deviation of 12.44, indicating a relatively diverse age distribution.
- The average tenure of customers is 31.26 months, with some customers having very short tenures (min=1.0) and others with longer tenures (max=60.0).
- The usage frequency feature has a mean of 15.81 and a standard deviation of 8.59, indicating a varying usage pattern among customers.
- The total spends feature ranges from 100 to 1000, with an average spend of approximately 631.62 and a standard deviation of 240.8.
- The Churn feature is binary, representing

whether a customer churned or not. The mean value is 0.57, suggesting that around 57% of customers churned.

E. Scale Data

Scaling the data is an essential purpose in data preprocessing for data analysis and machine learning tasks. Scaling ensures that all features are on a similar scale, preventing any single feature from dominating the analysis due to its larger magnitude. This is crucial when features have different units or scales, as it ensures that all features contribute equally to the analysis.

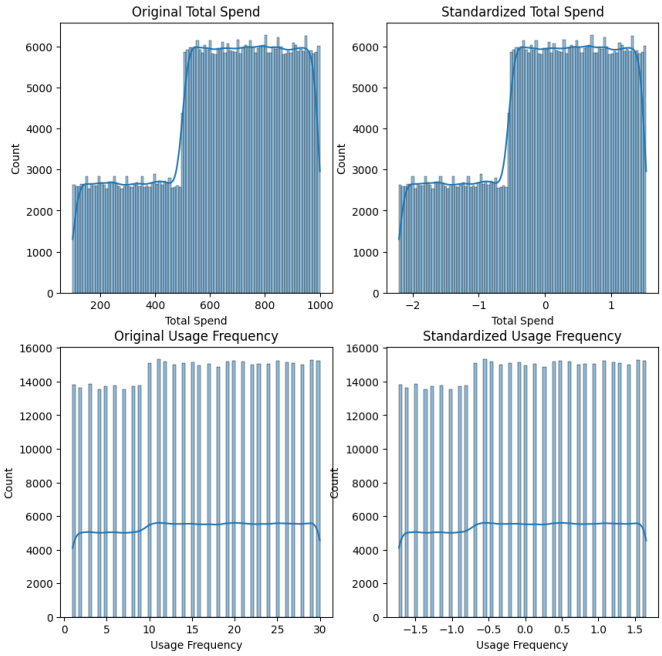

Fig. 7 Comparison of scaled data and normal data

III. K-Means Clustering

A. Elbow Method

The elbow method is a graphical technique used to determine the optimal number of clusters in a clustering algorithm, such as K-means clustering. The method helps to find the "elbow point," which is the point on the plot where adding more clusters does not significantly improve the clustering performance.

For this project, we chose all the regular numerical columns for getting the elbow point. The columns chosen are 'Age', 'Tenure', 'Usage Frequency', 'Support Calls', 'Payment Delay', 'Total Spend', 'Last Interaction.'

	Age	Tenure	Usage Frequency	Support Calls	Payment Delay	Total Spend	Last Interaction
0	-0.753326	0.448760	-0.210511	0.454549	0.609620	1.247427	0.293052
1	2.059646	1.028278	-1.724562	2.083100	-0.601319	-0.309865	-0.986584
2	1.255940	-1.000036	-1.375166	0.780259	0.609620	-1.854698	-1.335575
3	1.497051	0.390808	0.604748	1.105969	-0.722413	-0.978462	1.689018
4	-1.315921	0.043097	0.488282	0.454549	-0.601319	-0.060698	0.642043

Fig. 8 First 5 rows of the scaled and selected data for K-Means

We use a loop through different numbers of clusters, from 1 to 9, and performed K-means clustering with a specific number of clusters. For each clustering result, we calculated the sum of squared distances (inertia) of data points to their respective cluster centers and stored the inertia values in a 'result' list. After the loop, we plotted a elbow curve by using these values.

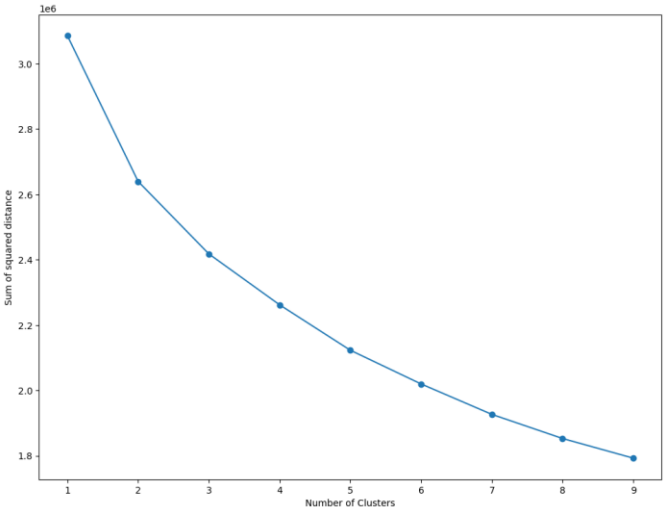

Fig. 9 Elbow Curve

After getting the curve, we used the 'KneeLocator' function from the 'kneed' library to find the elbow point on the previous elbow curve. We found out the Elbow point is 3 after running 'KneeLocator'.

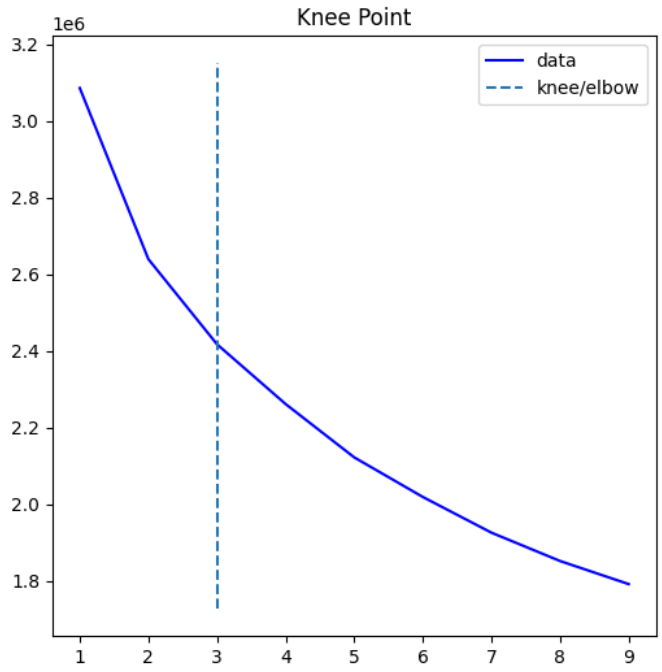

Fig. 10 Elbow point found in the curve.

Running the 'KMeans' by 3 clusters and add another column 'Cluster' to show each customer belonging to which cluster. We have a count plot to show the distribution of customers among the

clusters.

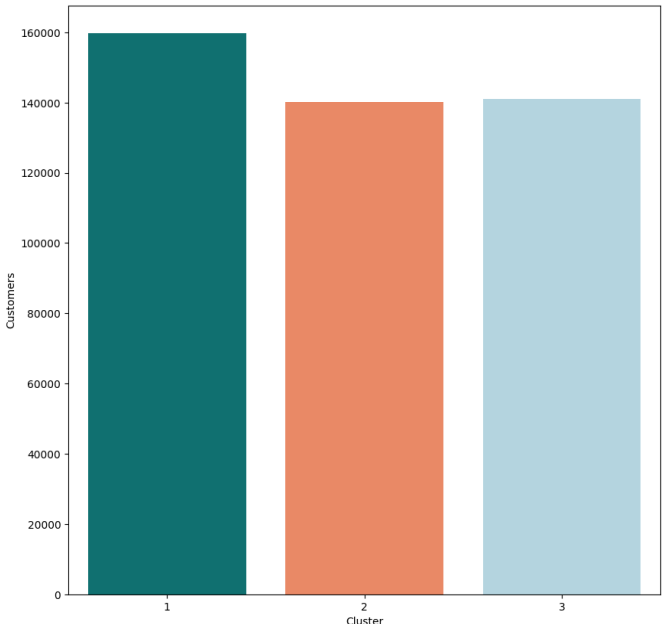

Fig. 11 Each cluster customers.

B. Analyze Clusters

By grouping the data by cluster and getting the means of all the clusters.

Cluster	Age	Tenure	Usage Frequency	Support Calls	Payment Delay	Total Spend	Last Interaction
1	36.746407	45.658118	14.253468	2.132727	10.714136	719.870708	13.504230
2	37.118285	15.420134	18.278688	2.294828	10.970953	713.461492	13.667735
3	44.589313	30.687535	15.111051	6.572916	17.498468	450.312768	16.395176

Fig. 12 Table for 3 clusters' mean.

We can get some information from this table.

- The age in cluster 3 is obviously higher than cluster 1 and 2.
- Cluster 1 has the highest average tenure, suggesting that customers in this cluster have been with the company for a longer duration compared to customers in the other clusters.
- Cluster 2 has higher usage frequency implying that customers in this cluster use the company's services more frequently than in the other two clusters.
- Cluster 3 needs three times customer support than the other two clusters.
- Cluster 3 has the highest payment delay, indicating that customers in this cluster have more frequent delays in making payments.
- Cluster 3 has very low total spend compared to other clusters.
- The last interaction for cluster 3 is higher than the other two clusters.

This information can give a brief description of each cluster of customers. The heatmap provides a visual presentation of the mean values of different features in 3 clusters.

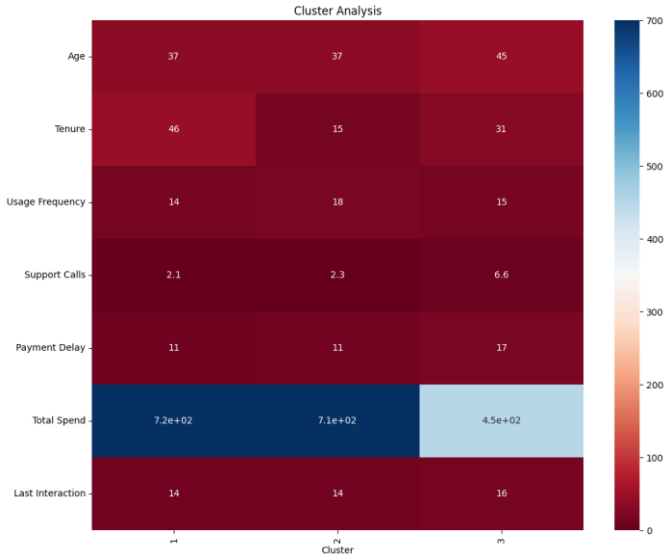

Fig. 13 Heatmap for 3 clusters' mean.

By analyzing the cluster means heatmap, businesses and data analysts can gain insights into the specific attributes or behaviors that define each customer segment. These insights can be leveraged to tailor marketing strategies, product offerings, and customer service approaches to meet the unique needs and preferences of each cluster. Additionally, it can help in making data-driven decisions and formulating targeted actions to maximize customer satisfaction, retention, and engagement across different segments.

We can also see the customer churn for each cluster.

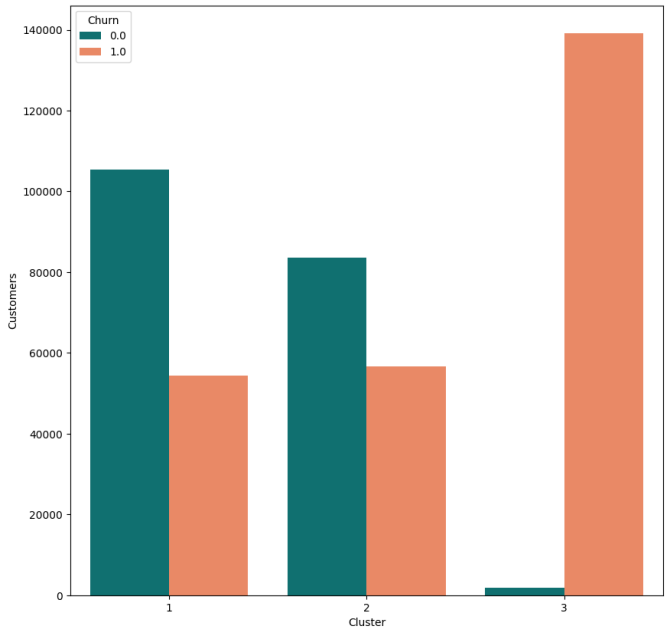

Fig. 14 Customer churn for 3 clusters.

From the graph we can see that almost all customers in cluster 3 are churned from the company. It indicates that there is a strong association between the characteristics of customers in Cluster 3 and their likelihood to churn. Understanding the reasons behind this high churn rate in Cluster 3 is crucial for the business, as it can provide valuable insights into customer behavior and satisfaction levels.

IV. PREDICTION

A. Separate Train and Test Data

For the train and test data, we drop the 'CustomerID' and 'Cluster' column to leave just useful values. We separate the data into X and Y. X contains the customer-related features. Y contains the binary labels indicating whether each customer churned (1) or not (0).

We use 'train_test_split' to separate X and Y into train data and test data (X_train, X_test, Y_train, Y_test). The shape of test and train data is in Fig 15.

```
X_train shape:  (352665, 10)
X_test shape:  (88167, 10)
Y_train shape:  (352665,)
Y_test shape:  (88167,)
```

Fig. 15 Shape of test and train data.

B. Imbalanced Data

We count the churn value shown in graph.

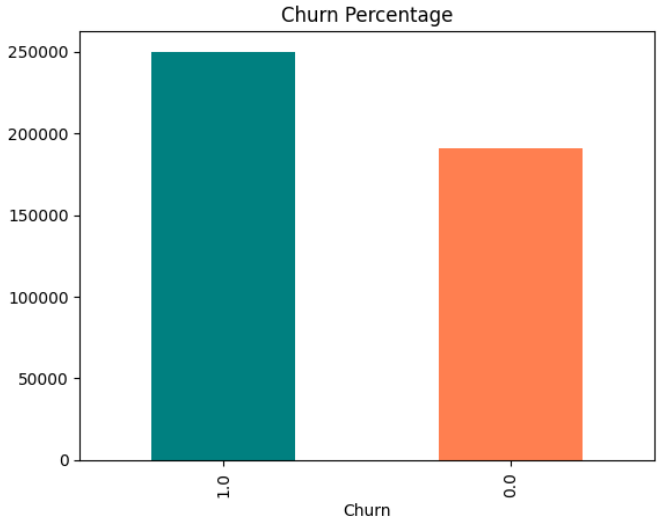

Fig. 15 Imbalanced Data

We can see from the graph the churn data is imbalanced. The imbalance can lead to biased predictions where the model trends favor the majority class and perform poorly on the minority class. The data needs to be oversampling.

The 'RandomOverSampler' library is used to make the minority class to have the same amount of data as the majority class by replicating the random selected instances.

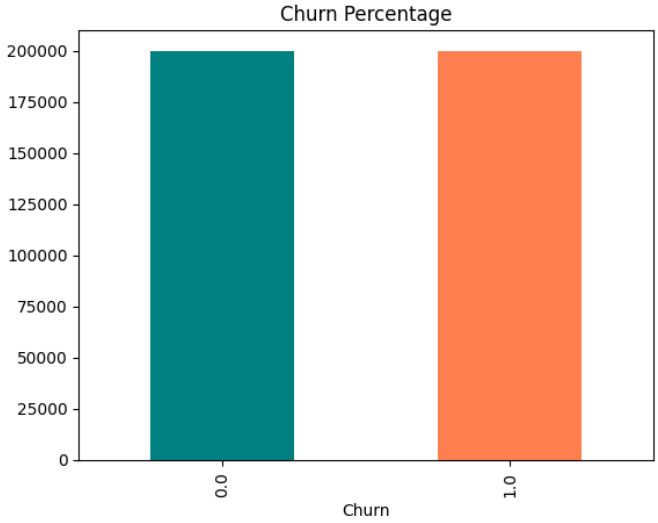

Fig. 16 Data after oversampling.

C. Feature Importance

Feature importance helps us to understand the relevance and contribution of each feature in the dataset towards predicting the target variable. The feature importance values indicate how much each feature affects the model's predictions. Higher values suggest that a feature has a stronger impact on the target variable, while lower values imply that the feature has less influence.

We use 'RandomForestClassifier' to get the importance of each feature in Fig.17.

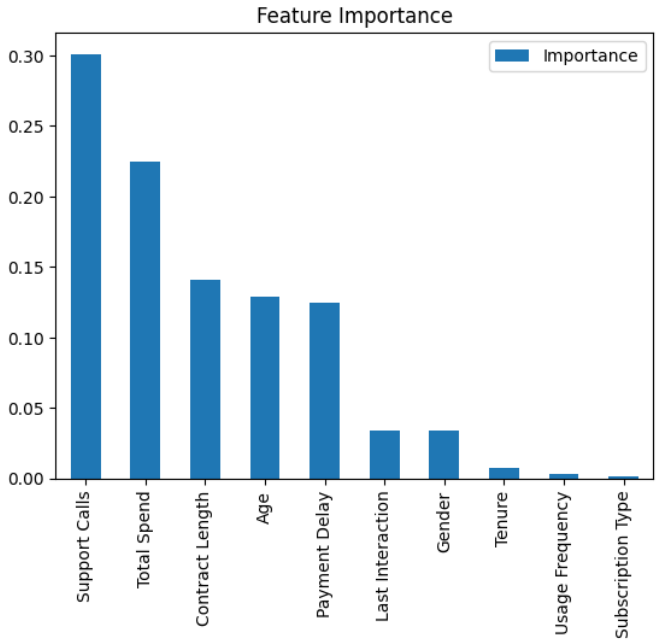

Fig. 17 Importance of features.

From the feature importance graph shown, we decided to use the first seven columns to predict the

churn.

D. Select Model

The churn value in this project is a binary classification problem (1 for churned and 0 for not churned). There are four models to consider solving the binary classification problem. We evaluate all of them with the precision and time cost of the model for consideration.

- Logistic Regression

```
Accuracy:  0.8700874476845079
Classification Report:
              precision    recall  f1-score   support

         0.0       0.82      0.90      0.86     38063
         1.0       0.91      0.85      0.88     50104

    accuracy                           0.87     88167
   macro avg       0.87      0.87      0.87     88167
weighted avg       0.87      0.87      0.87     88167

Time taken:  0.3170015811920166
```

Fig. 18 Logistic Regression Result.

- Decision Tree

```
Accuracy:  0.9940907595812493
Classification Report:
              precision    recall  f1-score   support

         0.0       0.99      0.99      0.99     38063
         1.0       0.99      1.00      0.99     50104

    accuracy                           0.99     88167
   macro avg       0.99      0.99      0.99     88167
weighted avg       0.99      0.99      0.99     88167

Time cost:  1.269829273223877
```

Fig. 19 Decision Tree Result.

- Random Forest

```
Accuracy:  0.9974253405469167
Classification Report:
              precision    recall  f1-score   support

         0.0       0.99      1.00      1.00     38063
         1.0       1.00      1.00      1.00     50104

    accuracy                           1.00     88167
   macro avg       1.00      1.00      1.00     88167
weighted avg       1.00      1.00      1.00     88167

Time cost:  30.709083557128906
```

Fig. 20 Random Forest Result.

- Neural Network

```
Accuracy:  0.9972325246407386
Classification Report:
              precision    recall  f1-score   support

         0.0       0.99      1.00      1.00     38063
         1.0       1.00      1.00      1.00     50104

    accuracy                           1.00     88167
   macro avg       1.00      1.00      1.00     88167
weighted avg       1.00      1.00      1.00     88167

Time cost:  35.29682731628418
```

Fig. 21 Neural Network Result

Based on the accuracy and the run time of the four model's results. We decided to choose the decision tree model.

E. Decision Tree

A Decision Tree is a popular machine learning algorithm used for both classification and regression tasks. It is a tree-like structure where each internal node represents a test on a feature, each branch represents the outcome of the test, and each leaf node represents a class label (in the case of classification) or a numerical value (in the case of regression).

After fitting the training data into the 'DecisionTreeClassifier' and use 'tree' library to get the visualized tree for the model, we got an extremely complicated tree and not readable.

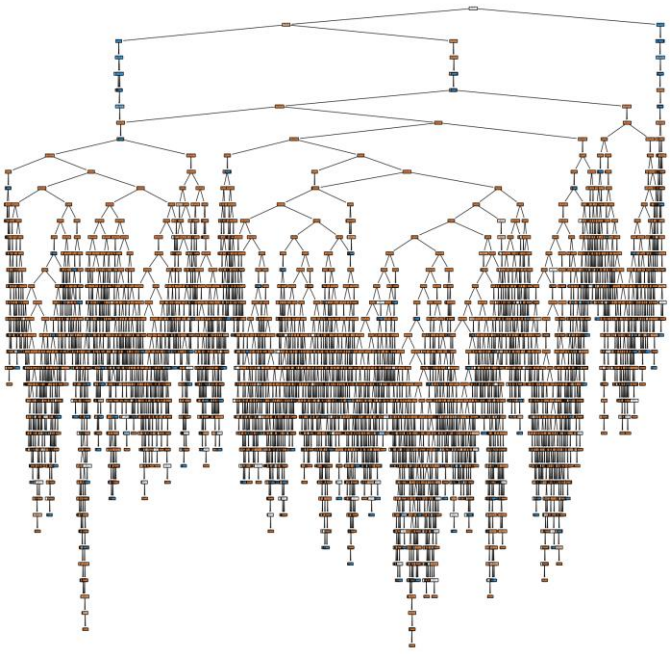

Fig. 22 Complicated Decision Tree.

For making a more readable tree, we decided to modify the max depth and max leaf nodes for the model. But the max depth and max leaf nodes are not easy to choose. The accuracy of the model should not be changed when max depth and max leaf nodes are

applied to the model. We made a for loop to find the best max depth and max leaf nodes. After 9 minutes computing, we got a table of records and sorted it with accuracy and sorted the values by max depth and max leaf nodes.

	max_depth	max_leaf_node	accuracy
336	11.0	28.0	0.997425
337	11.0	29.0	0.997425
338	11.0	30.0	0.997425
339	11.0	31.0	0.997425
340	11.0	32.0	0.997425
...	...	...	...
987	32.0	28.0	0.997425
988	32.0	29.0	0.997425
989	32.0	30.0	0.997425
990	32.0	31.0	0.997425
991	32.0	32.0	0.997425

110 rows × 3 columns

Fig. 23 Table for sorted values.

From the table on the top, we applied the DecisionTreeClassifier to have max depth to be 11 and max leaf nodes to be 28. After using the new model, the decision tree is more readable. A clearer figure for the decision tree can be seen in Appendix 1.

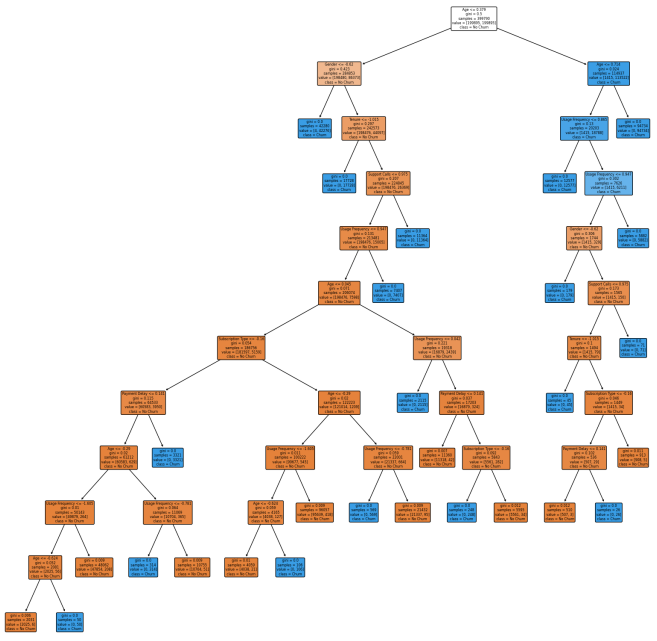

Fig. 24 Readable Decision Tree

V. CONCLUSION

In conclusion, this report focused on Customer Segmentation Analysis and Customer Churn Prediction using a dataset containing customer demographics, subscription details, and behavior. The analysis involved data preprocessing, including data cleaning, managing missing values, and converting non-numeric data into numerical form. Correlation analysis helped understand the relationships between different numerical variables. K-Means clustering was applied to segment customers into distinct groups, uncovering valuable insights about each cluster.

The analysis also addressed the issue of imbalanced data in customer churn prediction and employed RandomOverSampler to balance the data for more accurate predictions. Feature importance was examined using the RandomForestClassifier, identifying the most relevant features for churn prediction.

Among several models, the Decision Tree emerged as the chosen model due to its interpretability and accuracy. The decision tree provided a readable representation, allowing for easy understanding of the factors influencing customer churn.

Overall, the insights gained from this analysis empower businesses to improve customer engagement, enhance retention strategies, and tailor marketing approaches based on the unique characteristics of each customer segment. By leveraging data-driven strategies, businesses can foster long-lasting customer loyalty, reduce churn rates, and ultimately drive growth and success in their operations.

VI. REFERENCES

- Wikimedia Foundation. (2023, June 19). K-means clustering. Wikipedia. https://en.wikipedia.org/wiki/K-means_clustering
- Azeem, M. S. (2023, June 14). Customer churn dataset. Kaggle. https://www.kaggle.com/datasets/muhammadshahidazeem/customer-churn-dataset
- K-means clustering algorithm - javatpoint. www.javatpoint.com. (n.d.). https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning
- Patel, H. (2021, September 2). What is feature engineering - importance, tools and techniques for machine learning. Medium. https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10

VII. APPENDIX

1. Readable Decision Tree