Churn Analysis Project

Madan Thevar

Objective

The primary objective of this project is to analyze customer churn and predict which customers are likely to leave the company. By identifying these customers, the company can proactively retain them, thereby reducing churn rates.

Data Cleaning

The raw data required significant cleaning to make it suitable for analysis. This included handling missing values, removing duplicates, and filtering out irrelevant columns.

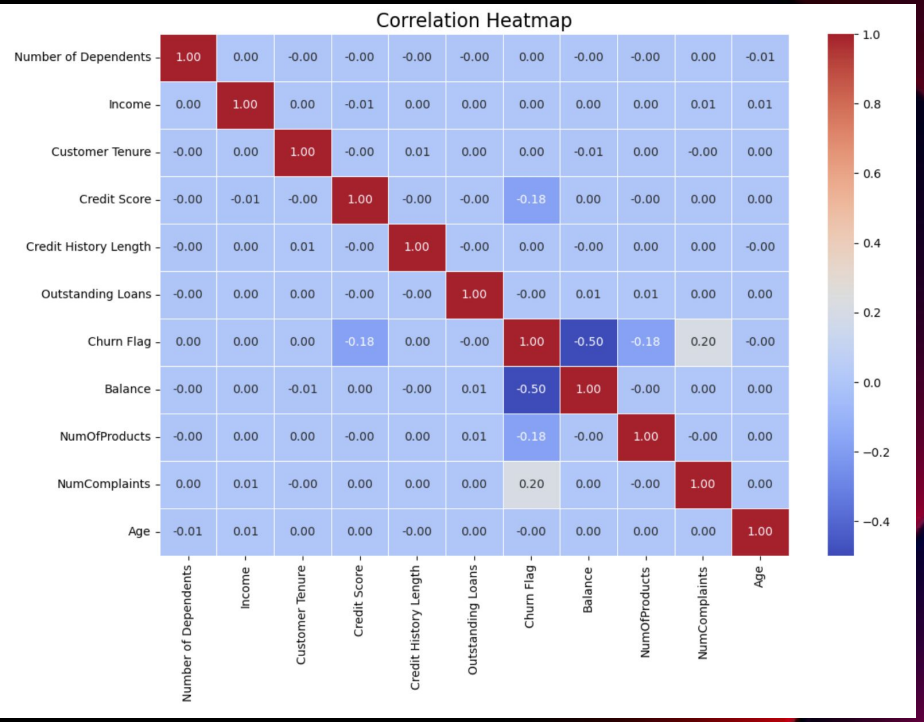
Specific attention was given to transforming the data into formats that are most effective for model building, such as converting date of birth into age and categorizing income levels.

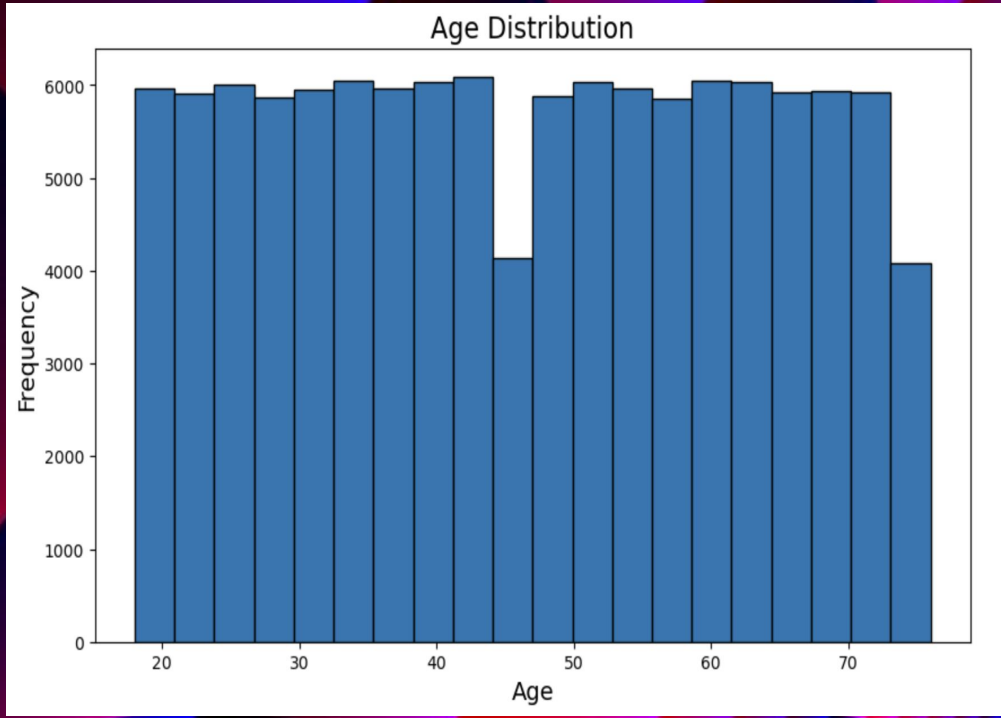
```
# Remove the 'RowNumber' column from the dataset
dataset_cleaned = dataset.drop(['RowNumber'], axis=1)

# Verify the removal
dataset_cleaned.head()
```

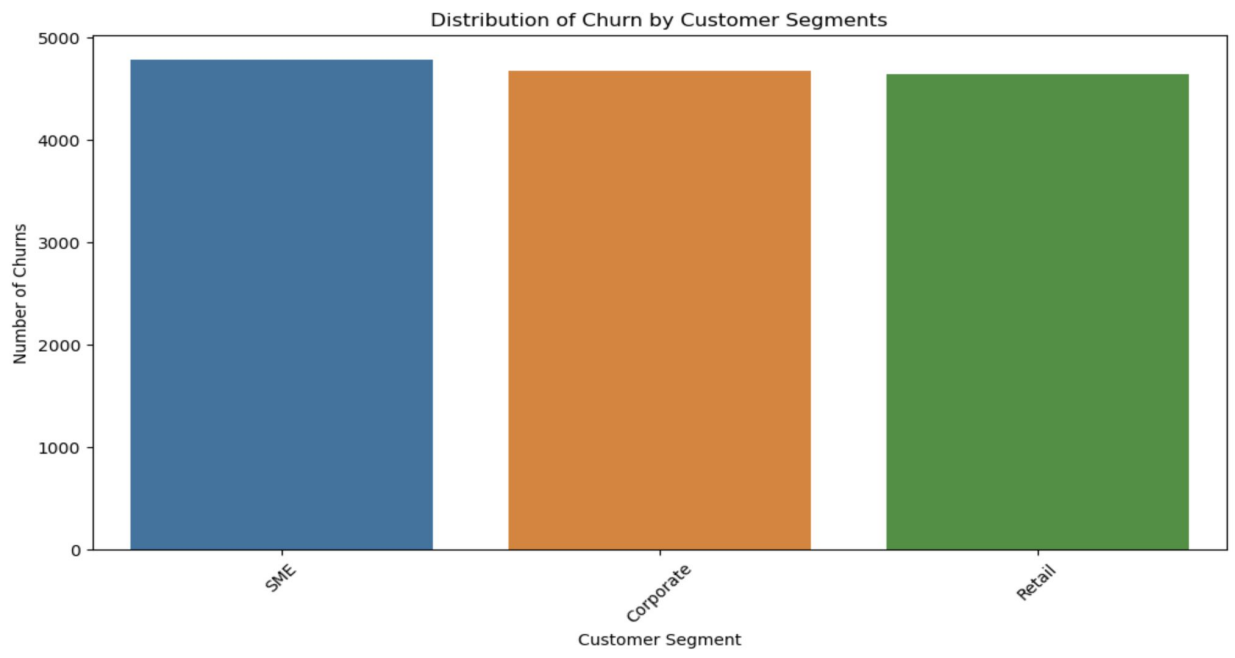
	CustomerId	Surname	First Name	Date of Birth	Gender	Marital Status	Number of Dependents	Occupation	Income	Education Level	...	Preferred Communication Channel	Credit Score	Credit History Length	Outstanding Loans	Churn Flag
0	83ef0b54-35f6-4f84-af58-5653ac0c0dc4	Smith	Troy	1987-08-29	Male	Divorced	3	Information systems manager	77710.14	High School	...	Phone	397	24	41959.74	0
1	009f115a-e5ca-4cf4-97d6-530140545e4e	Sullivan	Katrina	2000-02-07	Female	Married	1	Charity fundraiser	58209.87	High School	...	Email	665	10	8916.67	0
2	66309fd3-5009-44d3-a3f7-1657c869d573	Fuller	Henry	1954-02-03	Female	Single	1	Television production assistant	9794.01	High School	...	Email	715	21	43270.54	0
3	b02a30df-1a5f-4087-8075-2a35432da641	Young	Antonio	1991-01-15	Female	Divorced	5	Agricultural engineer	15088.98	High School	...	Phone	747	17	17887.65	0
4	0d932e5b-bb3a-4104-8c83-f84270f7f2ea	Andersen	John	1992-04-08	Female	Divorced	2	Teacher, early years/pre	60726.56	Master's	...	Email	549	25	32686.84	0

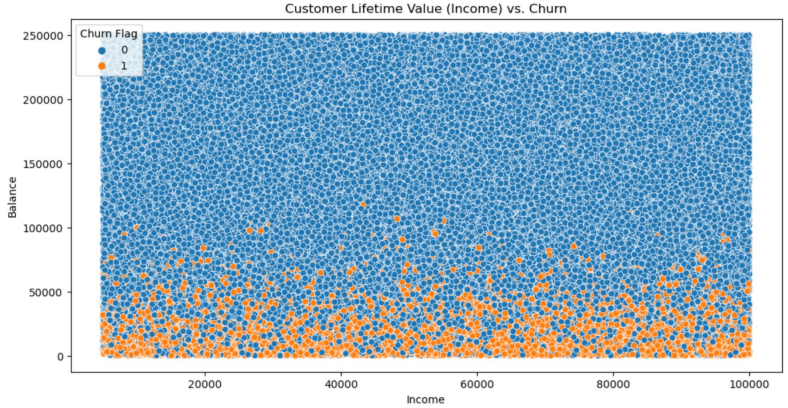
5 rows × 24 columns

Python Visualizations






Distribution of Churn by Customer Segments
Number of Churns
5000
4000
3000
2000
1000
0
SME
Corporate
Retail
Customer Segment


Customer Lifetime Value (Income) vs. Churn
Churn Flag
0
1
Balance
250000
200000
150000
100000
50000
0
20000
40000
60000
80000
100000
Income

Churn Prediction: Model Performance and Key Insights

Random Forest Classifier:
Accuracy: 0.9755274991352473

	precision	recall	f1-score	support
0	0.97	1.00	0.99	20301
1	0.98	0.81	0.89	2827
accuracy			0.98	23128
macro avg	0.98	0.91	0.94	23128
weighted avg	0.98	0.98	0.97	23128

Support Vector Machine (SVM):
Accuracy: 0.9948114839156001

	precision	recall	f1-score	support
0	1.00	1.00	1.00	20301
1	0.99	0.97	0.98	2827
accuracy			0.99	23128
macro avg	0.99	0.98	0.99	23128
weighted avg	0.99	0.99	0.99	23128

Decision Tree Classifier:
Accuracy: 0.9839156001383604

	precision	recall	f1-score	support
0	0.99	0.99	0.99	20301
1	0.94	0.93	0.93	2827
accuracy			0.98	23128
macro avg	0.97	0.96	0.96	23128
weighted avg	0.98	0.98	0.98	23128

Linear Regression (Logistic Regression):
Accuracy: 0.99805430646835

	precision	recall	f1-score	support
0	1.00	1.00	1.00	20301
1	1.00	0.99	0.99	2827
accuracy			1.00	23128
macro avg	1.00	0.99	1.00	23128
weighted avg	1.00	1.00	1.00	23128

Key Takeaways and Model Selection for the analysis

Support Vector Machine (SVM):

- **Accuracy**: 99.48%
- **Strength**: Best for high precision and minimal false positives.

Logistic Regression:

- **Accuracy**: 92.08%
- **Strength**: Good overall but lower recall for predicting churn.

Random Forest Classifier:

- **Accuracy**: 97.55%
- **Strength**: Balanced precision and recall, great for complex datasets.

Decision Tree Classifier:

- **Accuracy**: 98.39%
- **Strength**: High interpretability, helpful in understanding churn factors.

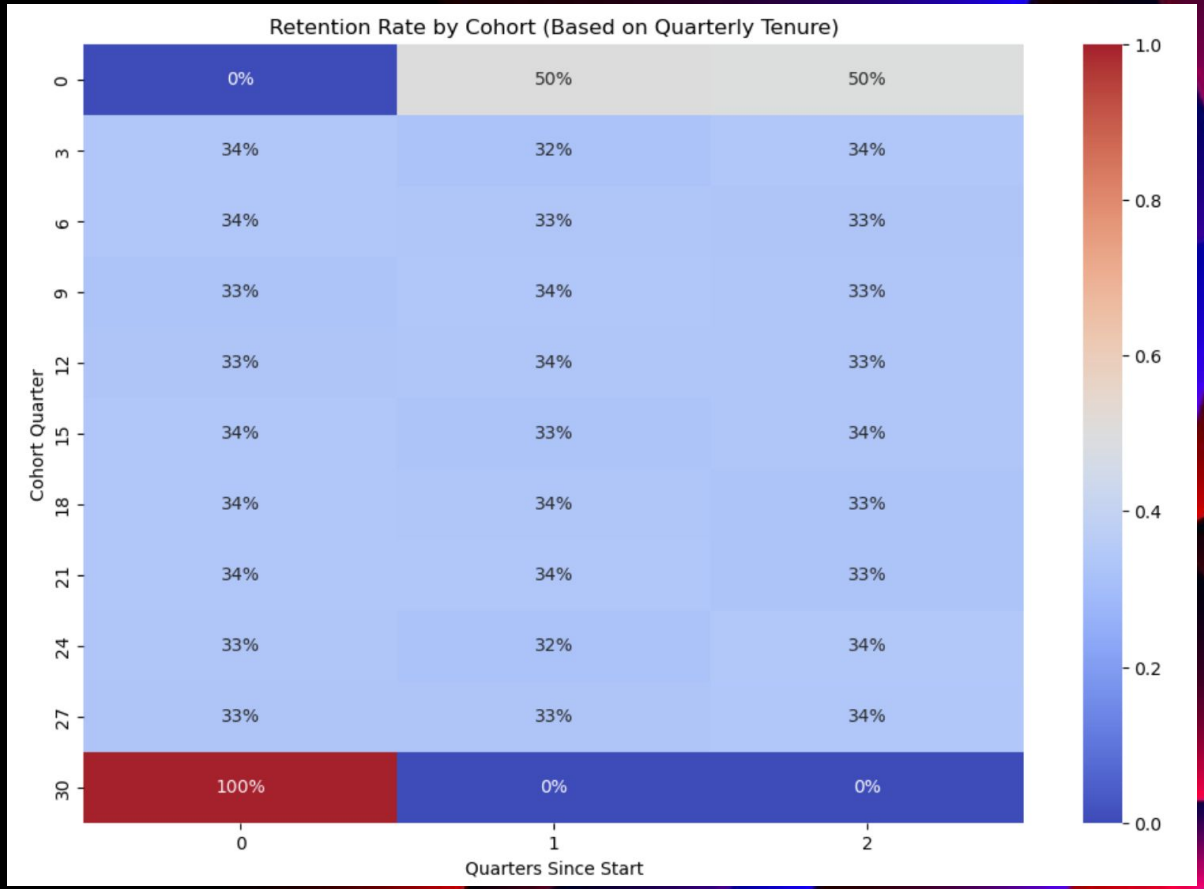
Linear Regression (Logistic Regression):

- **Accuracy**: 99.8%
- **Strength**: Nearly perfect accuracy, ideal for both churn and non-churn predictions.

Selection of the Models and their purpose

- **SVM and Logistic Regression**: These models should be the primary choices for deployment due to their high accuracy and balanced performance across metrics. SVM is powerful if maximizing accuracy is the top priority.
- **Random Forest**: A highly reliable model that balances performance and feature interpretability, suitable for deeper analysis and understanding of churn factors.
- **Decision Tree**: Ideal for situations where model interpretability is critical, and understanding the decision-making process behind churn prediction is essential.

Cohort Retention Analysis Heatmap - Unique Graph



Key Takeaways and Improving the analysis

1) **Early Drop-off:** Significant customer churn was observed in the first quarter, indicating a need for targeted retention strategies during the onboarding phase.
2) **Stable Retention:** Retention rates stabilize around 32-34% after the initial quarter across most cohorts, highlighting consistent customer behavior post-initial phase.
3) **Cohort Comparison:** Newer cohorts show lower retention rates than earlier ones, suggesting a potential decline in customer loyalty or changes in customer acquisition quality.
4) **Strategic Insights:** The heatmap identifies high-risk churn periods, allowing for data-driven strategies to enhance customer retention over time.
5) **Visualization Technique:** Utilized Seaborn's heatmap with quarterly cohort grouping and retention rates, providing a clear visual representation of retention patterns.
6) **Improving Retention:** By analyzing this graph, we can implement targeted interventions during high-risk periods, especially in the early stages, to decrease churn rates and enhance customer retention.

Tableau Churn Analysis Dashboard

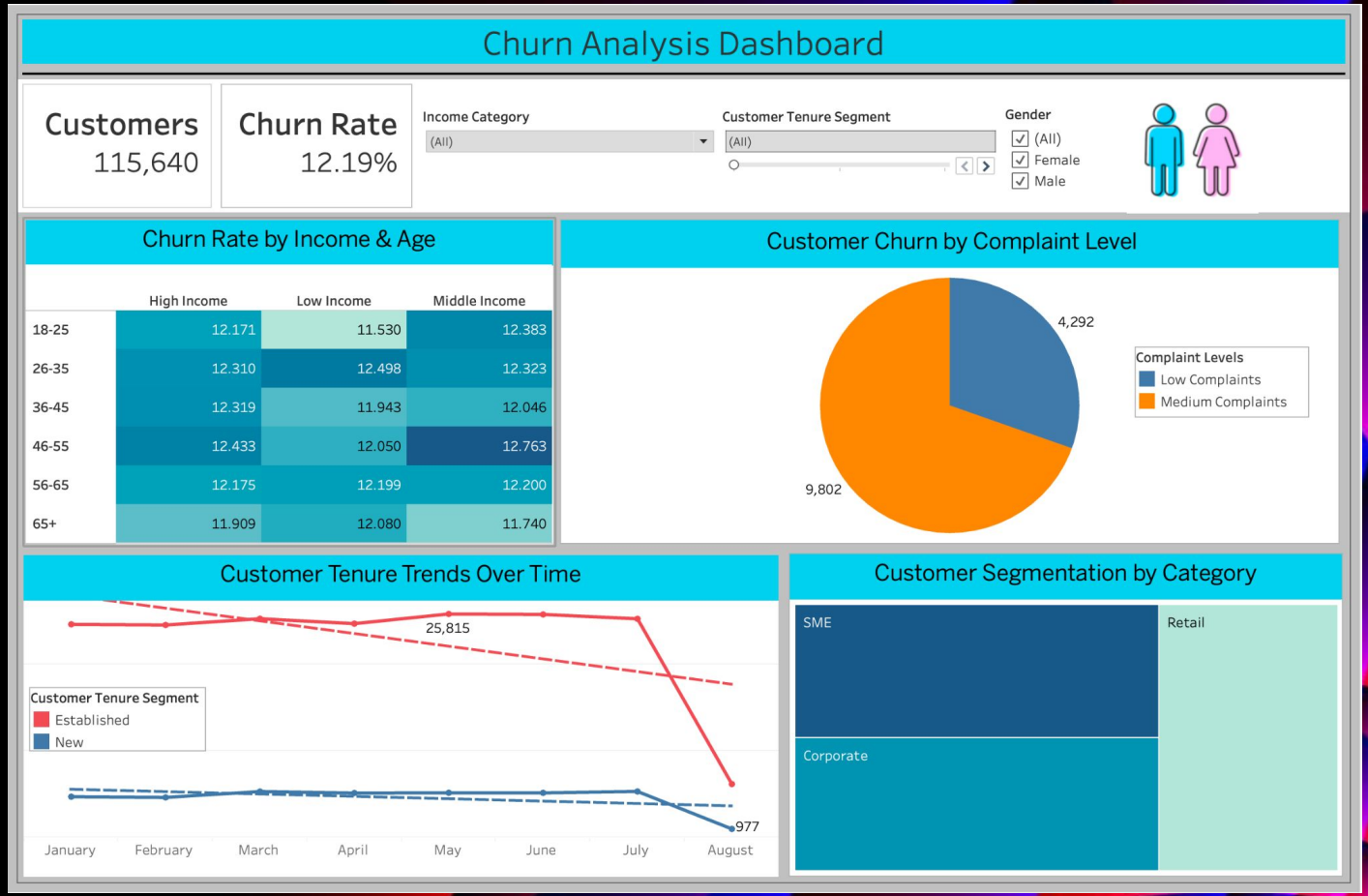


Tableau Dashboard Preparation

1. Data Preparation:

- **Calculated Fields:**
  - **Churn Rate:** As a percentage of total churned customers.
  - **Customer Age Group & Tenure Segment:** For detailed analysis.
- **Data Cleaning:**
  - Removed unnecessary columns, focusing on relevant data for the dashboard.

2. Strategic Recommendations:

- **Focus on High-Risk Segments:**
  - Implement retention strategies targeted at the 46-55 age group in the Middle Income category.
- **Enhance Customer Support:**
  - Address Medium Complaints proactively to mitigate churn.
- **Segmented Strategies:**
  - Utilize Small or Mid-size Enterprises' customer segmentation data (SME) to tailor engagement and retention efforts.

Improvement Suggestions

- **Focus on High-Risk Segments:** Consider implementing targeted retention campaigns for the **46-55 age group** within the **Middle-Income** category, as they show a higher propensity to churn.
- **Enhance Customer Support:** The correlation between complaint levels and churn indicates a need for improved customer service, particularly for those with **Medium Complaints**.

What I have learned from this Analysis?

This churn analysis has revealed critical insights, such as the varying churn risks across customer segments and the significant impact of early customer engagement. One unique technique that can be applied is **Cohort-based Predictive Analytics**. By tracking customer behavior in cohorts, we can tailor retention strategies to specific groups, focusing on the most vulnerable periods in the customer lifecycle. This approach, combined with machine learning models like SVM and Random Forest, allows us to predict churn more accurately and implement timely, personalized interventions.

Use cohort analysis to identify the most vulnerable customer groups and introduce cohort-specific retention initiatives. Regularly monitor the effectiveness of these strategies through the cohort heatmap, adjusting tactics as needed. By implementing these data-driven strategies, we can significantly reduce churn rates, increase customer lifetime value, and ultimately drive higher profitability for the organization. This proactive approach will mitigate revenue loss and strengthen our brand’s reputation for customer-centricity and innovation.