Customer Churn Prediction and Analytics System

Prepared by: **Shashank Chauhan**

Tools Used: **Python, MySQL, Power BI**

Introduction

This project aims to analyze customer behavior, segment users based on purchasing patterns, and predict customer churn in an online retail business. With customer acquisition costs rising, companies are increasingly focused on retaining existing customers and reducing churn. Utilizing a combination of data processing, statistical analysis, machine learning, and dashboard visualization, this project provides insights and predictions to inform customer retention strategies.

Technologies used include:

- **SQL** for data preprocessing and aggregation
- **Python** for data science, clustering, and machine learning
- **Power BI** for interactive dashboard visualization

Project Workflow Overview

1. Raw transaction data preprocessing and cleaning
2. RFM metric calculation and clustering-based customer segmentation
3. Exploratory and advanced data analysis
4. Churn prediction using machine learning models
5. Business dashboard creation for stakeholder insights

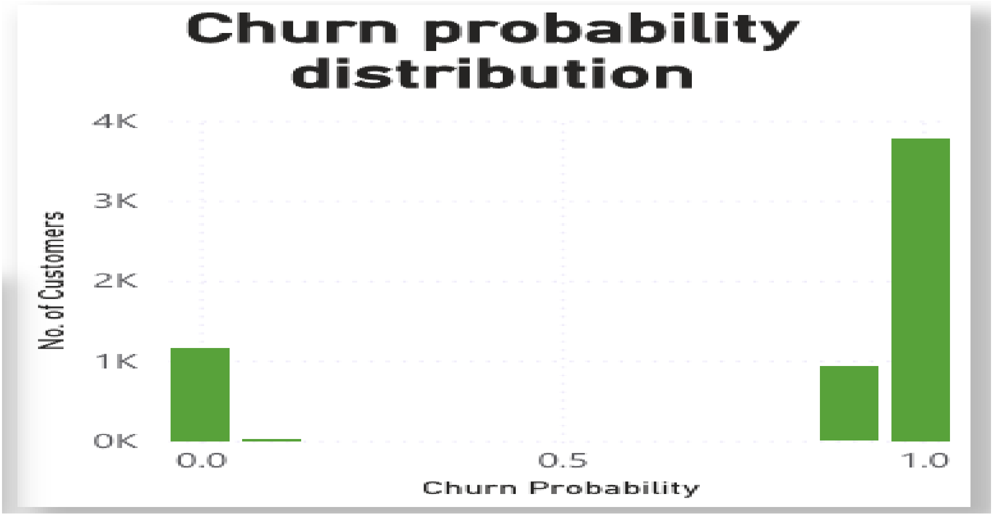
Key Results & Insights

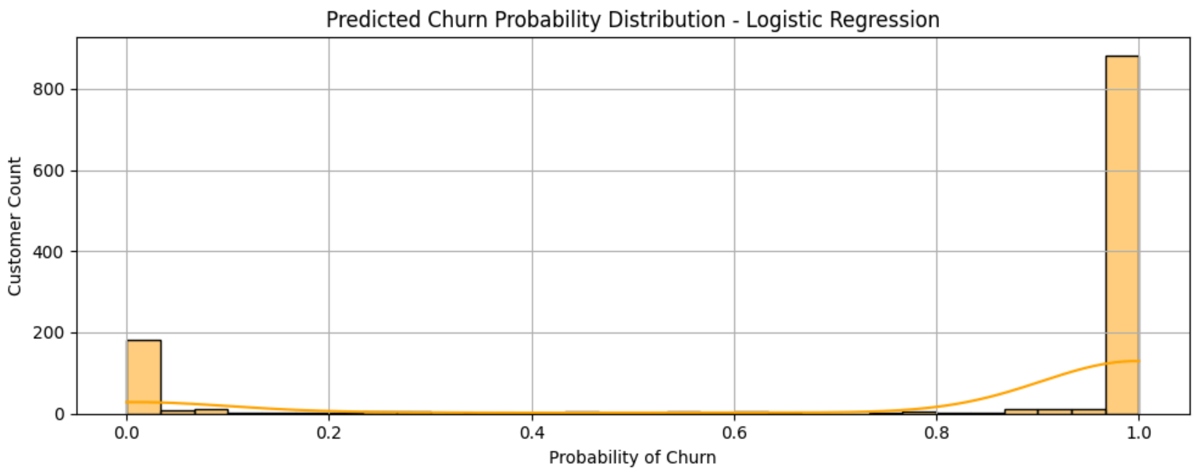
- Total customers analyzed: 5,860
- At-risk customers identified: 1,159 (≈ 19.78%)
- Average customer churn probability: 80.22%
- The highest churn was observed in segments: Potential Loyalists, VIP Customers, and One-Time Buyers

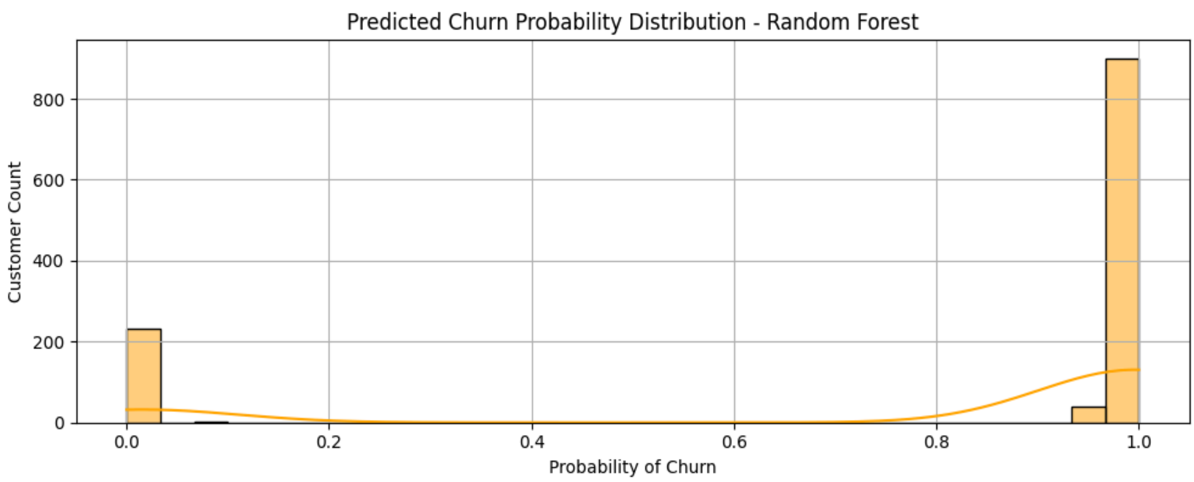
- Segment with lowest churn: At-Risk Customers (due to classification bias)
- Country-level and product-level churn variation visualized via a dashboard

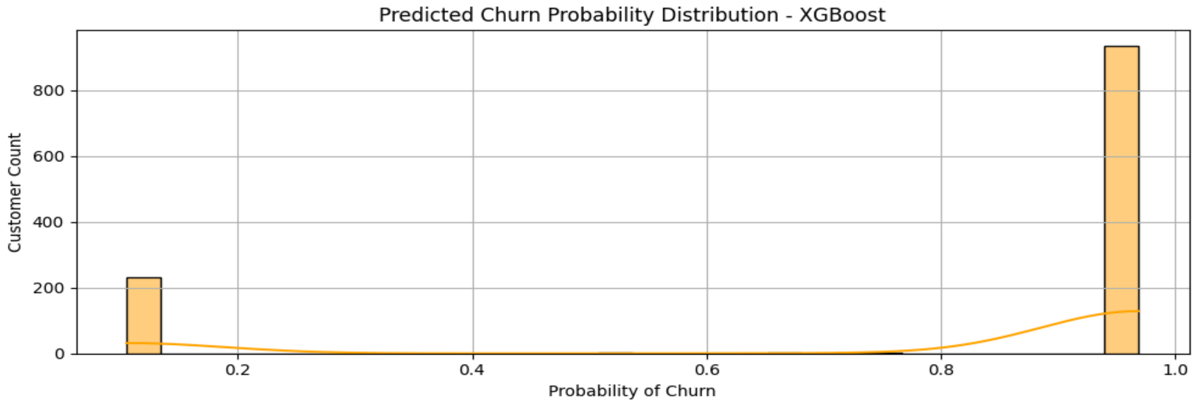
Visualizations:

- Churn probability distribution (Power BI/Python)

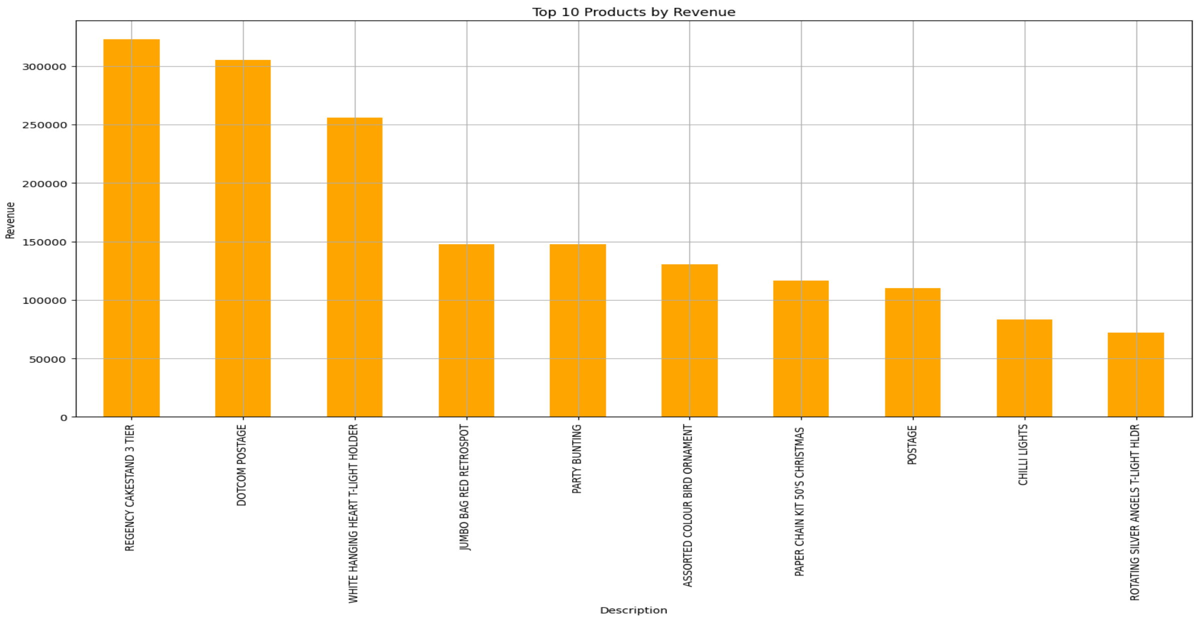


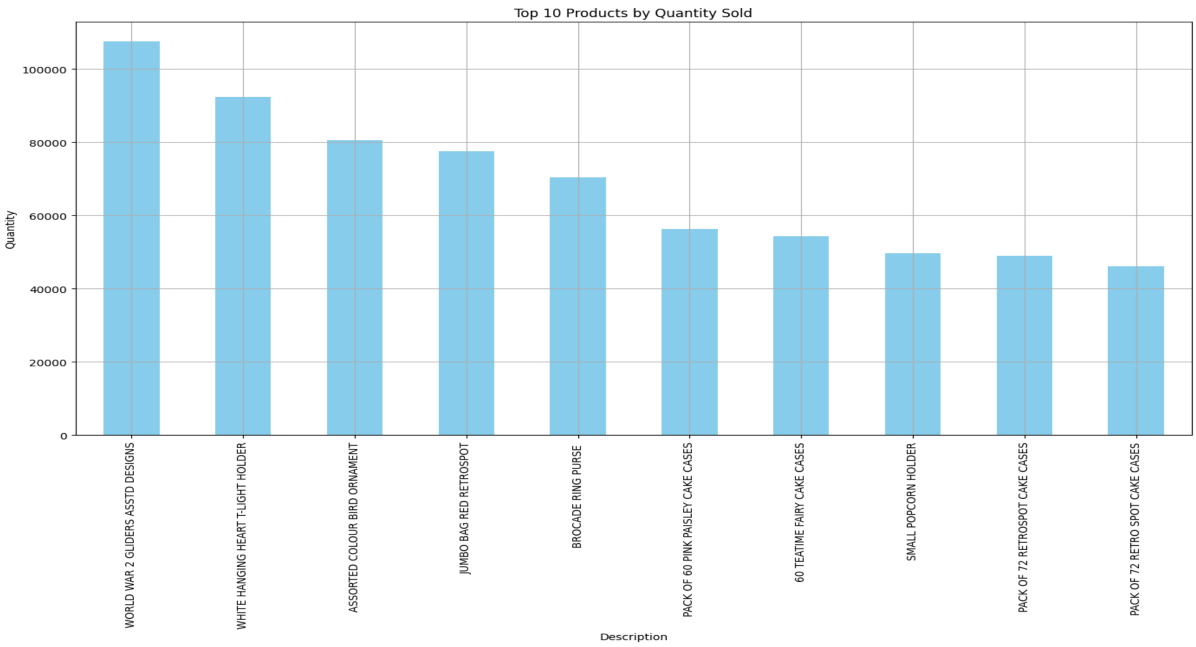




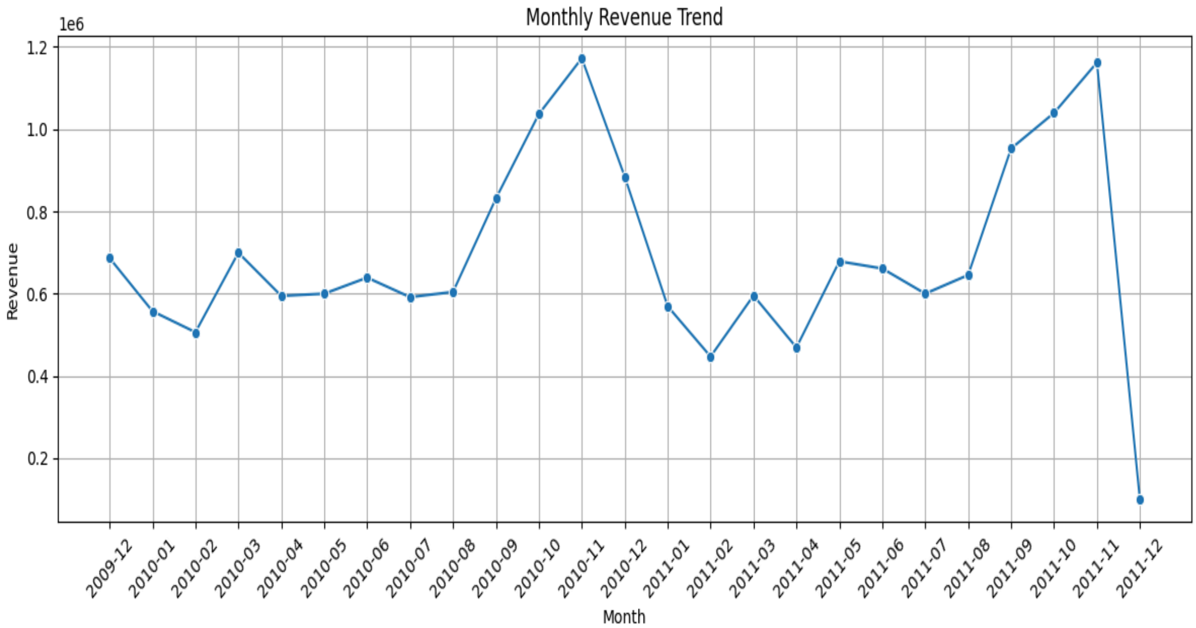


- Top-selling products (Python)



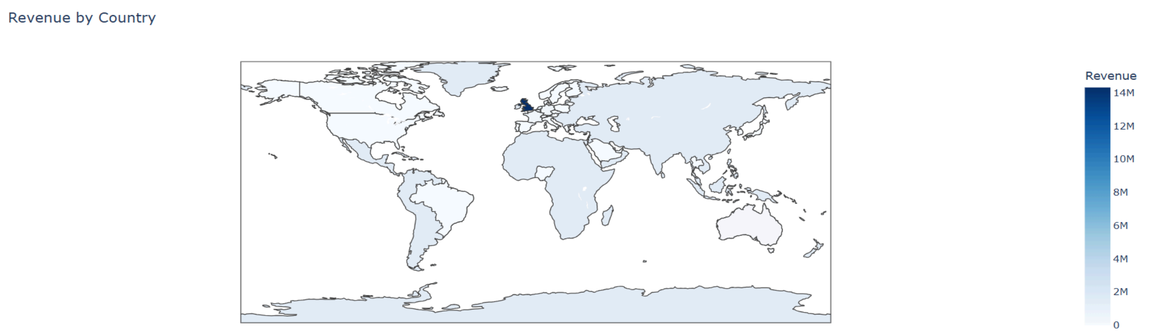


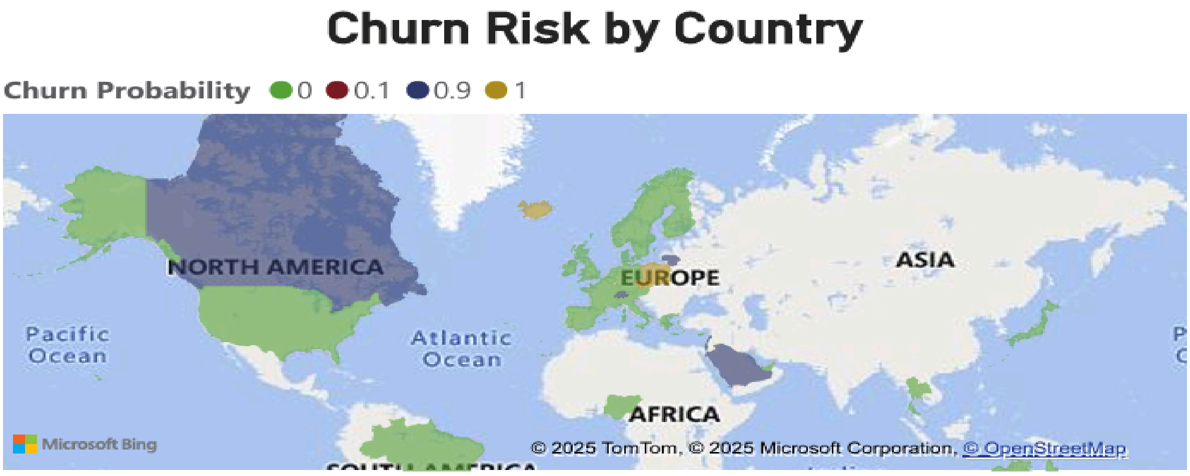
- Monthly Revenue trends (Python/SQL)



Month	Monthly_Revenue
2009-12	686654.16
2010-01	557319.06
2010-02	506371.07
2010-03	699608.99
2010-04	594609.19
2010-05	599985.79
2010-06	639066.58
2010-07	591636.74
2010-08	604242.65
2010-09	831615
2010-10	1036680
2010-11	1172336.04
2010-12	884591.89
2011-01	569445.04
2011-02	447137.35
2011-03	595500.76
2011-04	469200.36
2011-05	678594.56
2011-06	661213.69
2011-07	600091.01
2011-08	645343.9
2011-09	952838.38
2011-10	1039318.79
2011-11	1161817.38
2011-12	99713.7

- Country-wise churn risk (Python/Power BI)





Country	Total_Customers	Avg_Risk
Bahrain	2	1
Czech Republic	1	1
EIRE	5	1
European Community	1	1
Iceland	1	1
Israel	4	1
Malta	2	1
Poland	6	1
Singapore	1	1
Switzerland	22	1
Belgium	29	0.99
Canada	5	0.99
France	94	0.99
Australia	15	0.98
Germany	106	0.98
Norway	13	0.98
Lithuania	1	0.97
Portugal	24	0.97
Saudi Arabia	1	0.97
Spain	40	0.96
Finland	14	0.95
Lebanon	1	0.95
United Kingdom	5337	0.95
Unspecified	6	0.95
Netherlands	22	0.93

Customer Segmentation (RFM + Clustering)

Recency, Frequency, and Monetary (RFM) metrics were calculated for each customer:

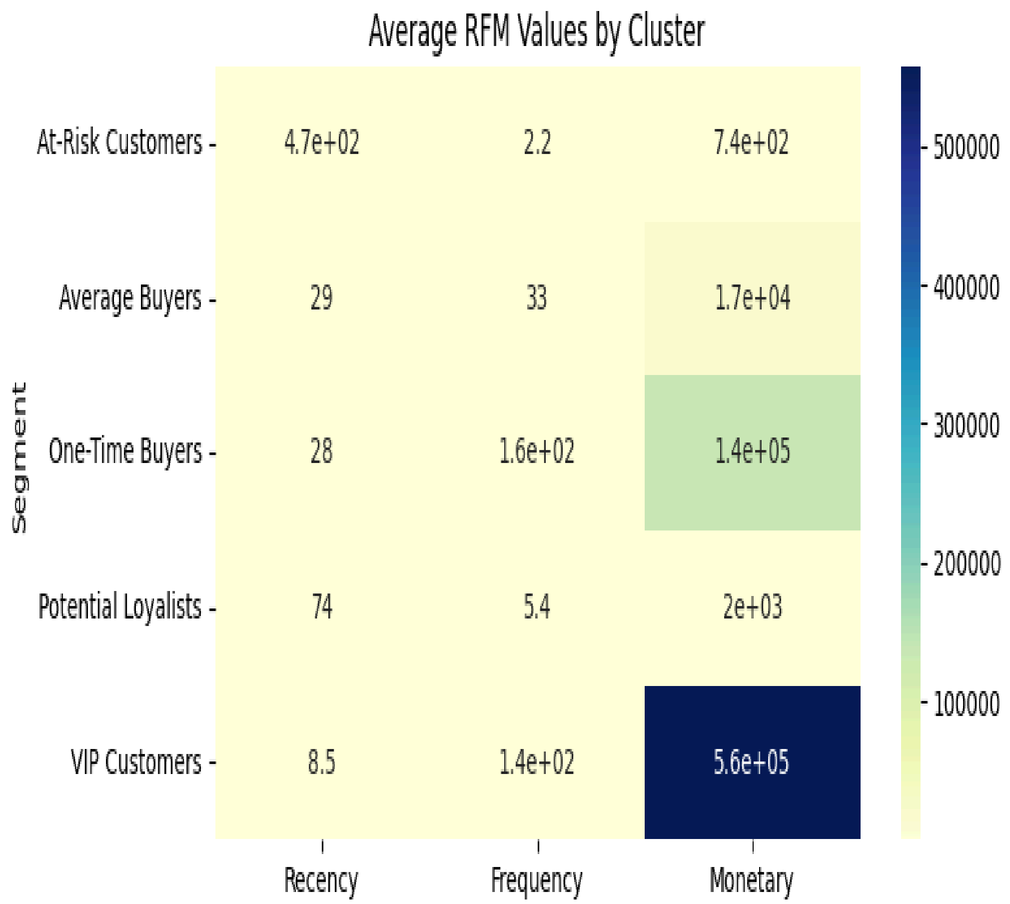
- Recency: Days since last purchase
- Frequency: Number of transactions
- Monetary: Total spend amount

Using K-Means clustering, customers were segmented into meaningful groups:

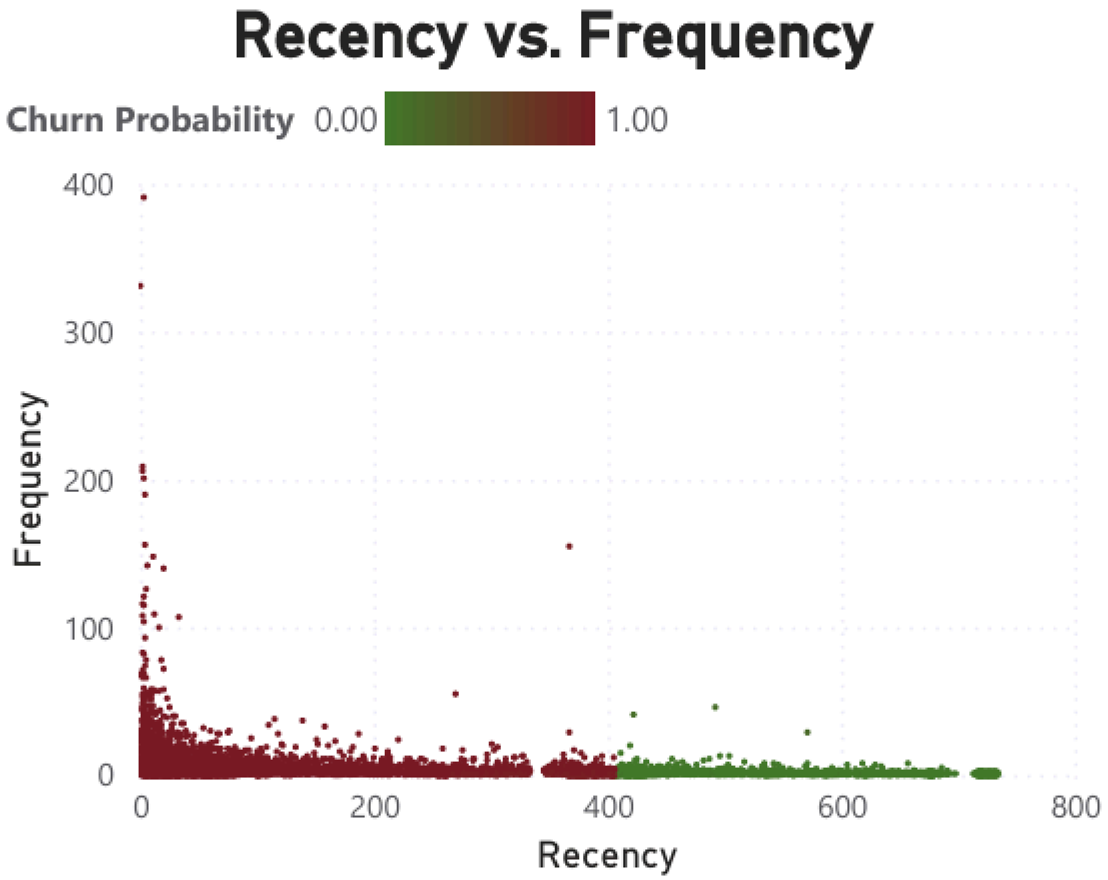
- VIP Customers
- Potential Loyalists
- Average Buyers
- One-Time Buyers
- At-Risk Customers

Visualizations:

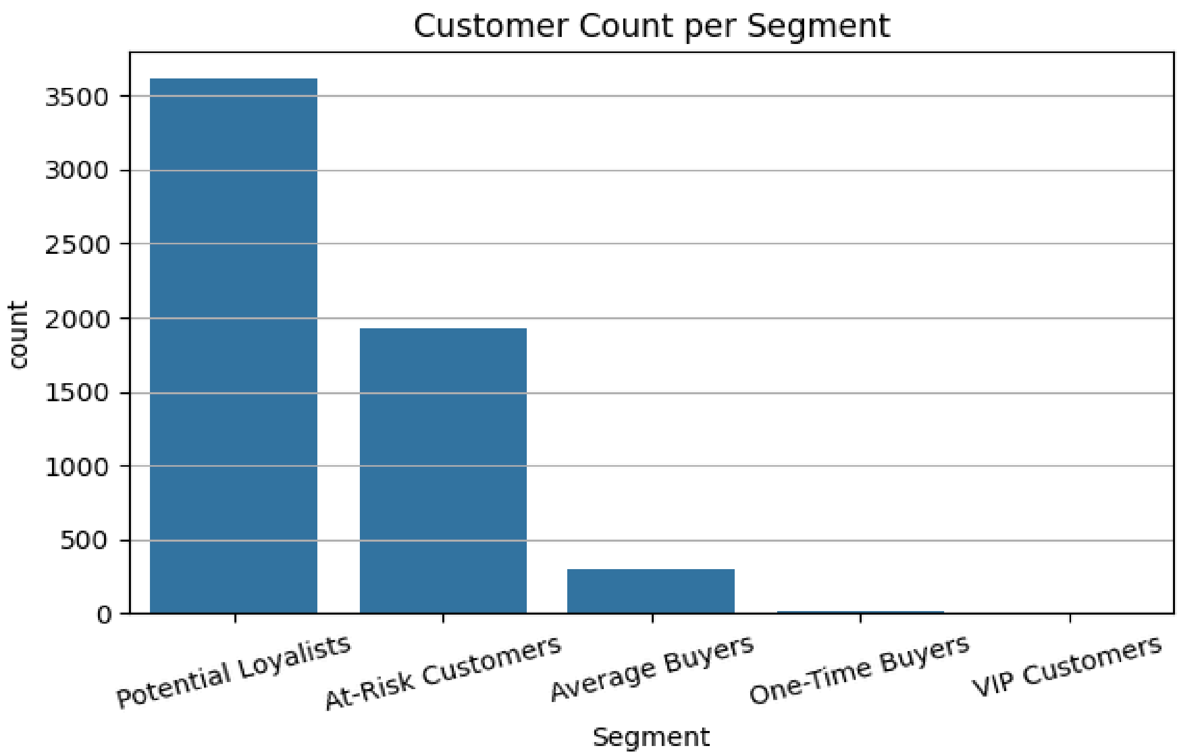
- RFM distribution plots (Python)



- Cluster scatterplots (Recency vs Frequency)



- Cluster distribution histogram/bar chart



Machine Learning for Churn Prediction

After preparing the labeled data with churn indicators, several models were trained and evaluated:

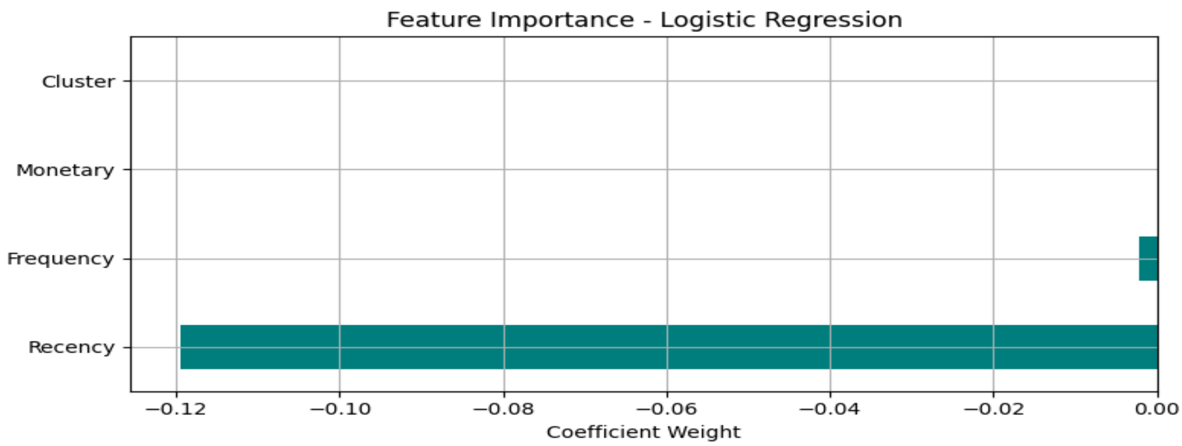
- Logistic Regression
- Random Forest (base and tuned)
- XGBoost Classifier

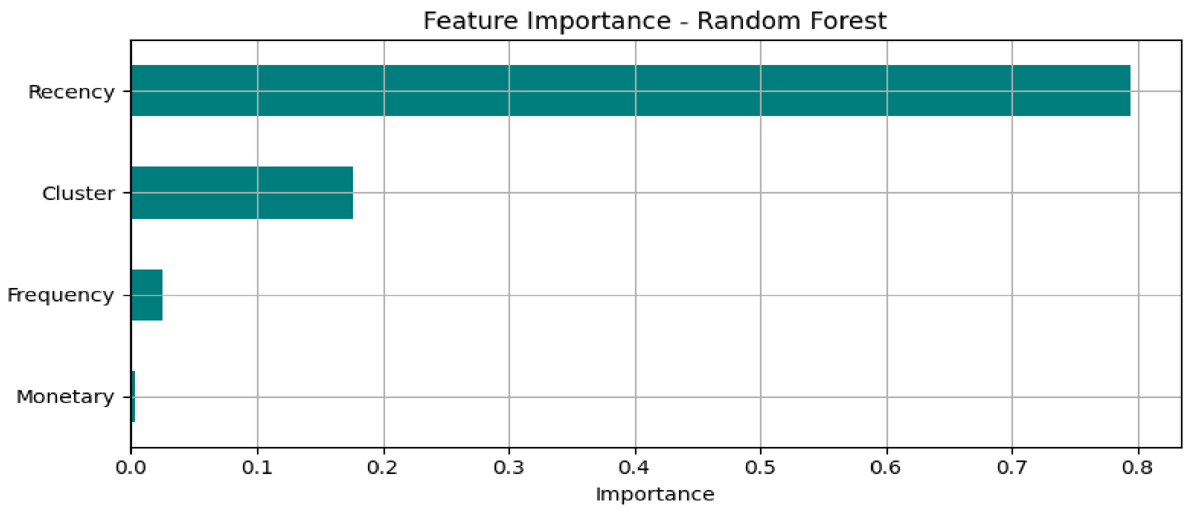
Model Evaluation:

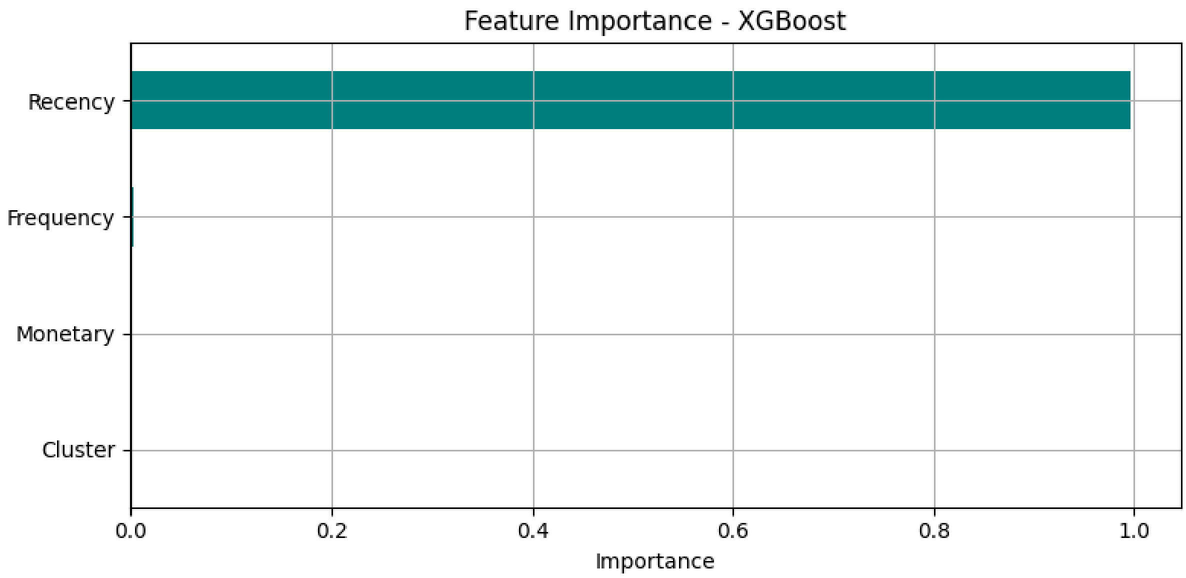
- Best performance achieved with the tuned Random Forest model
- Evaluation metrics considered: Accuracy, ROC-AUC, Precision, Recall
- Model output saved in churn_predictions.csv for business usage

Visualizations:

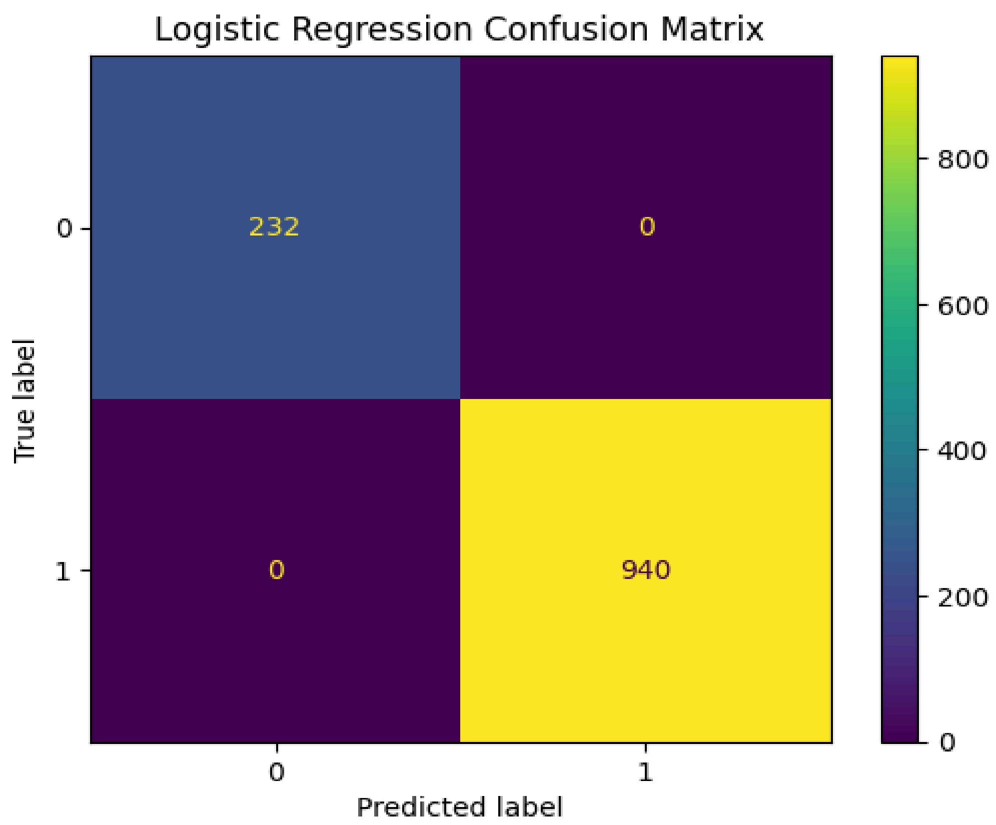
- Feature importance plot (Python)

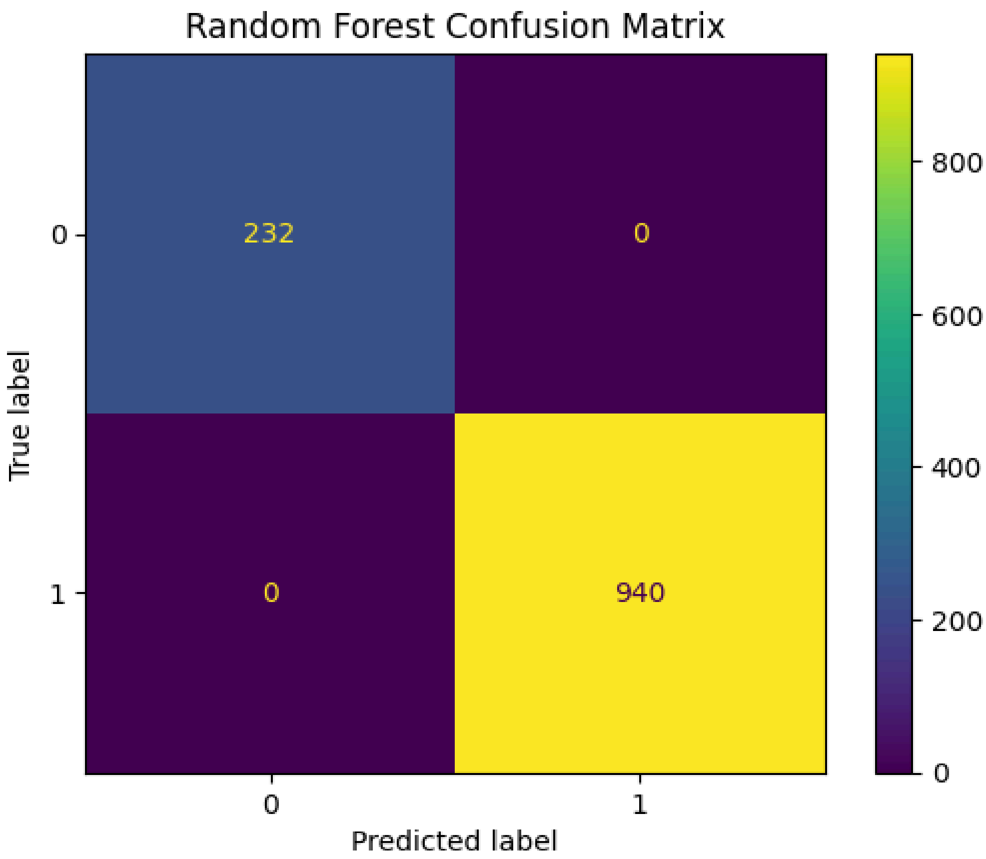


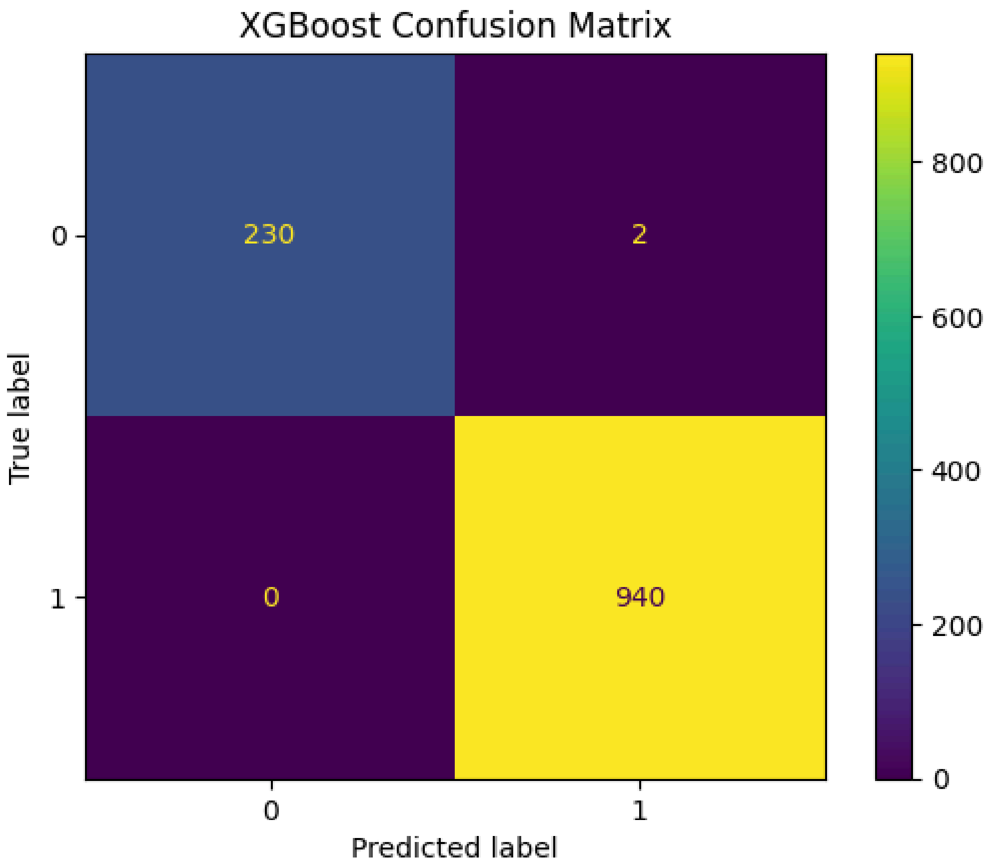




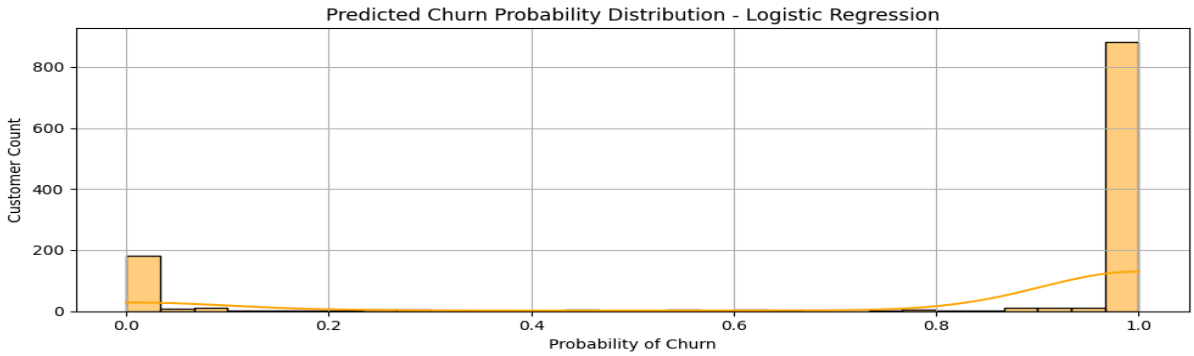
- Confusion matrix

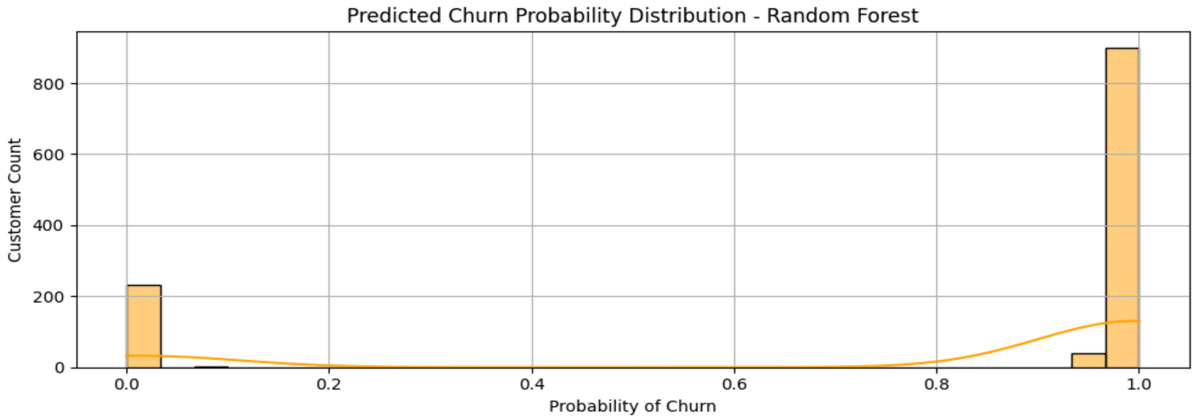


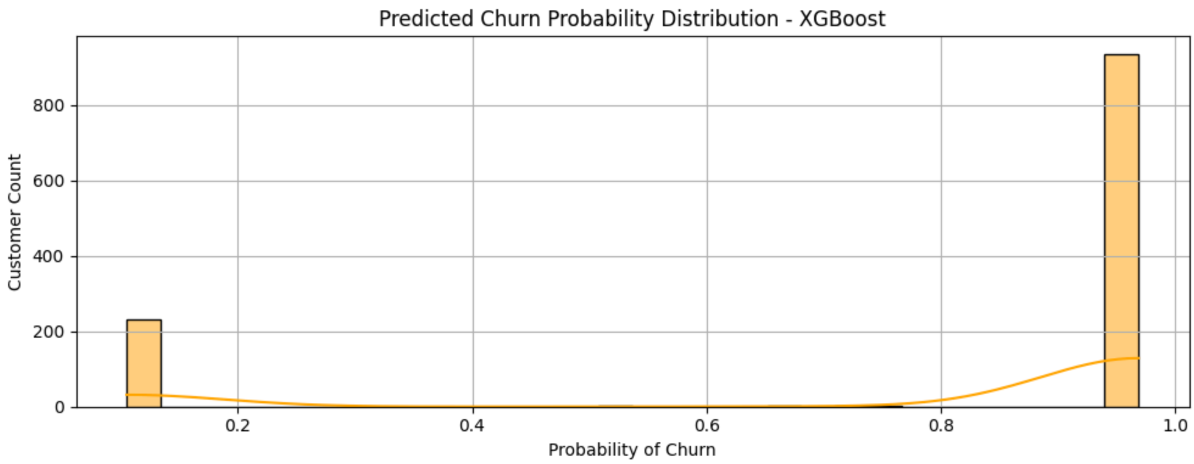

Random Forest Confusion Matrix
232
0
0
940
True label
Predicted label
0
1
0
1
800
600
400
200
0


XGBoost Confusion Matrix
230
2
0
940
True label
Predicted label
0
1
0
1
800
600
400
200
0

- Churn probability histogram







Power BI Dashboard Summary

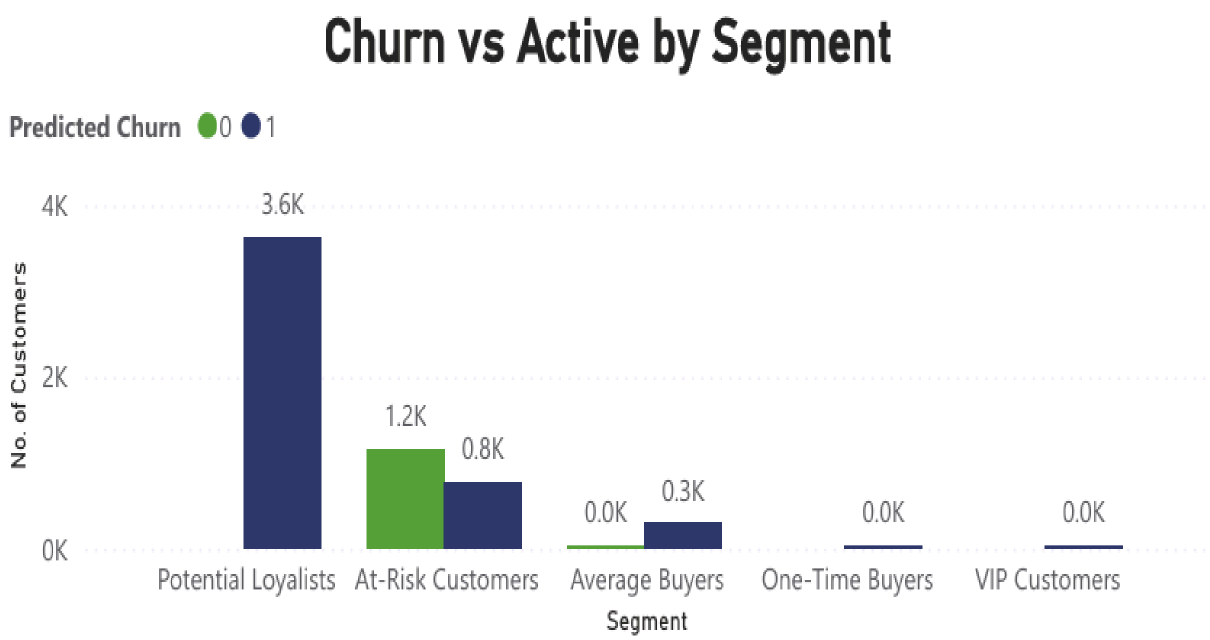
An interactive Power BI dashboard was created to present key findings:

- Overview of customer base and churn risk
- Visual segmentation of at-risk customers
- Churn risk distribution by RFM segment and country
- Monthly trends and key product insights

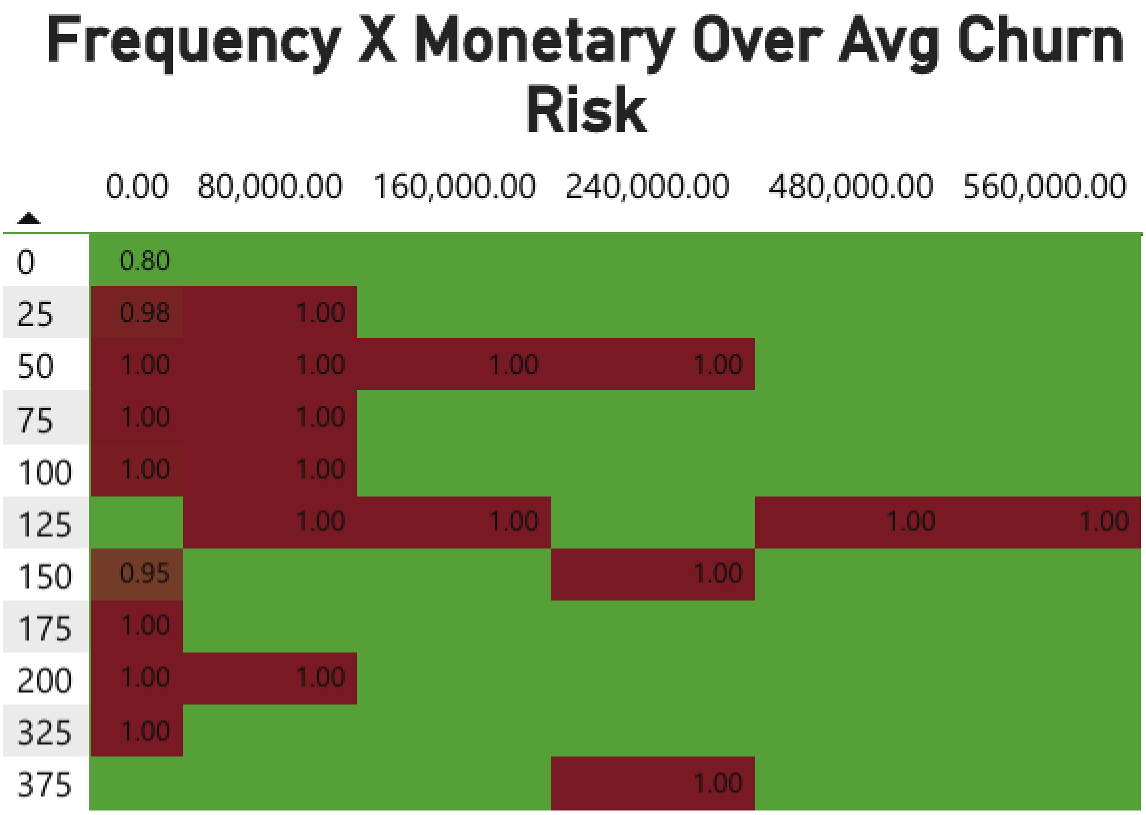
Visualizations:

- Dashboard summary panel

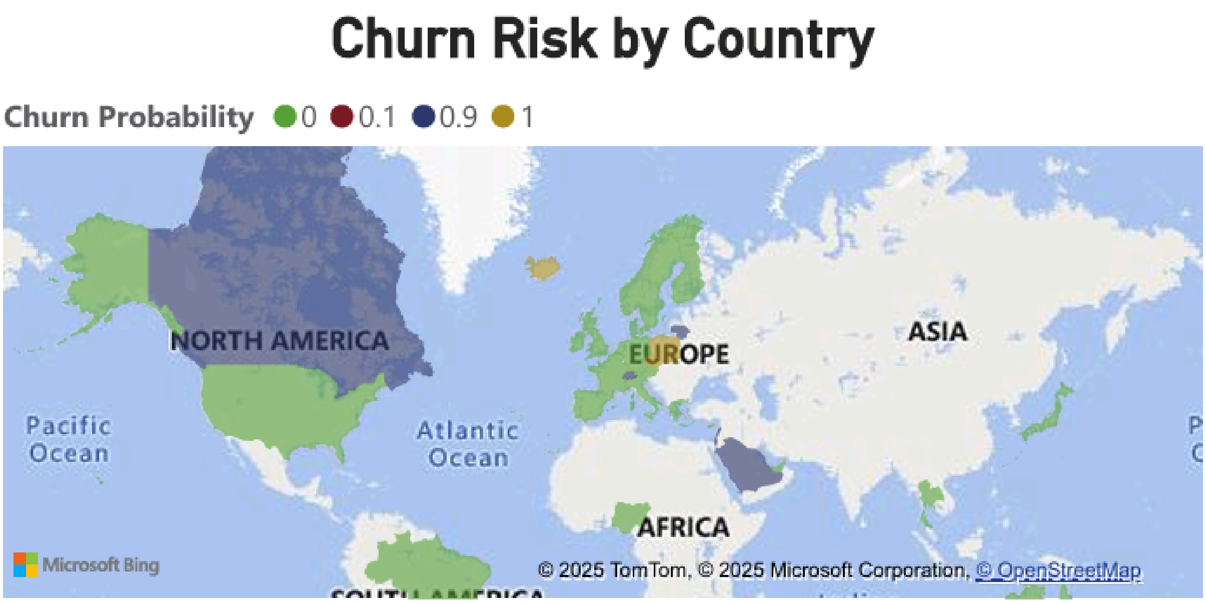
- Churn vs active by segment



- Recency vs Monetary heatmap



- Country-level churn map



Business Recommendations

Based on data-driven insights:

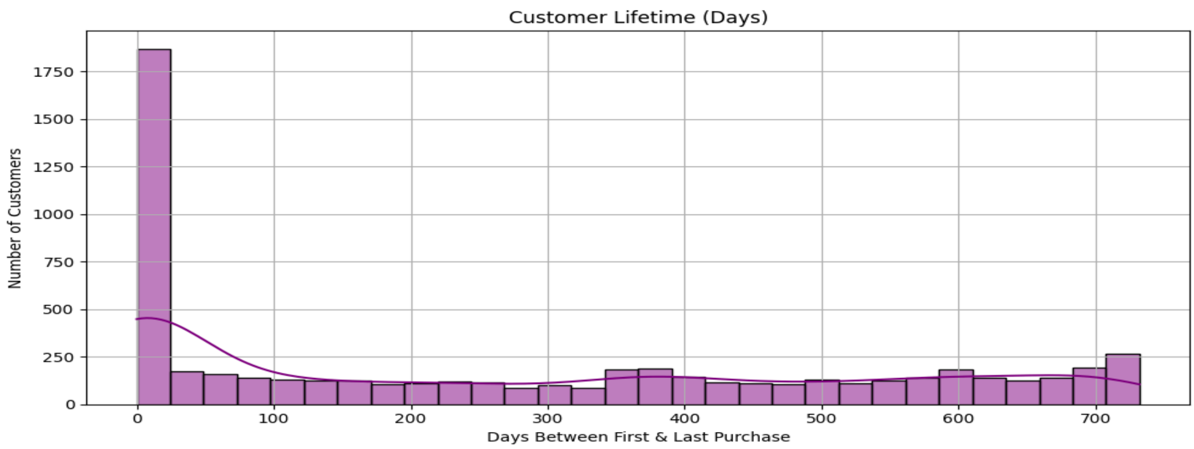
- Target Potential Loyalists and Average Buyers with Retention Offers
- Launch loyalty programs and incentives for high-value VIP Customers
- Monitor One-Time Buyers for re-engagement campaigns
- Implement CRM alerts for high-risk segments

Actionable suggestions:

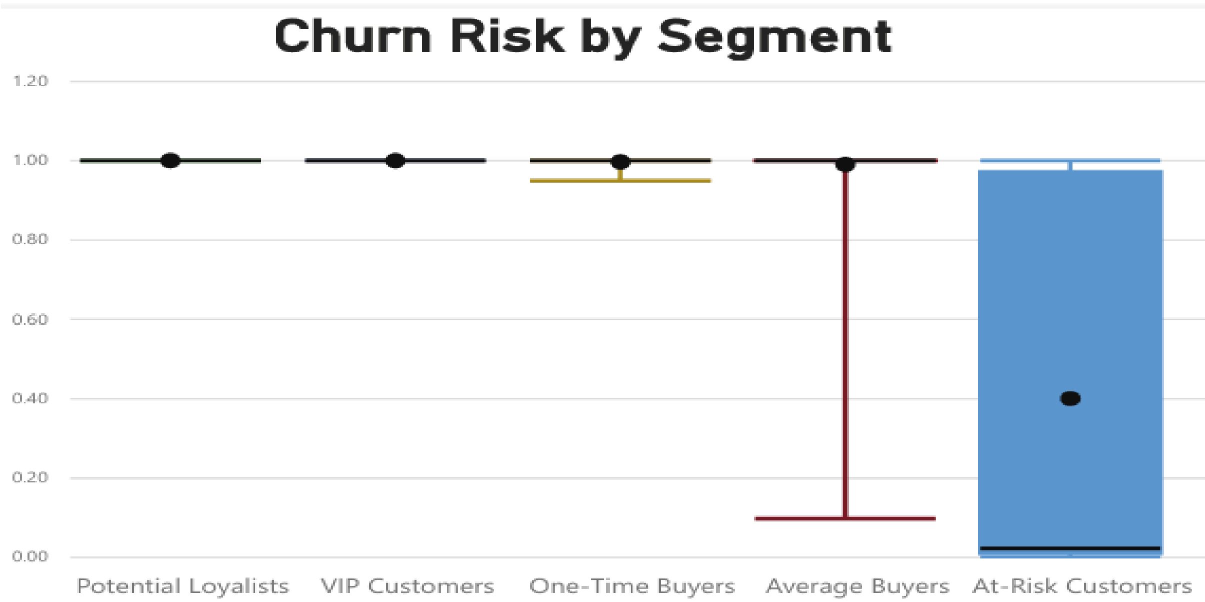
- Integrate churn probabilities into CRM workflows
- Personalize email campaigns based on segment behavior
- Focus retention budget on medium-risk, high-value customers

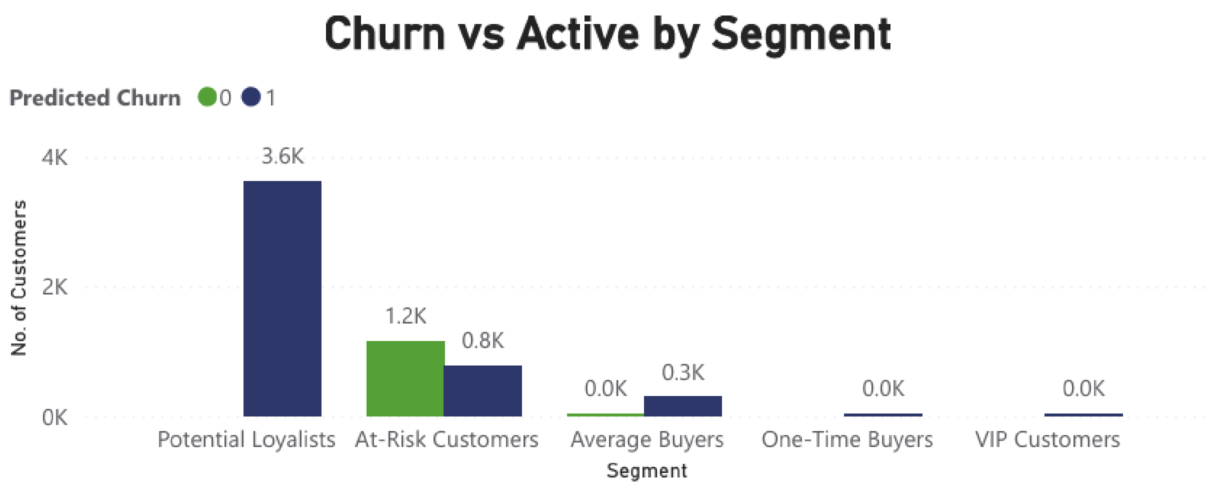
Visualizations:

- Retention curve (Python)



- Segment-level churn summary





SQL-Based Advanced Analytics

Several SQL views were created to support advanced business insights:

- Top_Stock_Revenue: Shows the top 10 products based on revenue generated. This view supports product-level analysis in identifying key SKUs.

```
CREATE VIEW Top_Stock_Revenue AS
SELECT Description AS Stock, ROUND(SUM(Quantity * Price), 2) AS Revenue
FROM customer_clean_transactions GROUP BY Description ORDER BY Revenue DESC LIMIT 10;

SELECT * FROM Top_Stock_Revenue;
```

- Monthly_Revenue_Trend: Extracts monthly aggregated sales revenue, enabling time-based analysis of business performance.

```
CREATE VIEW Monthly_Revenue_Trend AS
SELECT CONCAT(SUBSTRING(InvoiceDate, 7, 4), '-', SUBSTRING(InvoiceDate, 4, 2)) AS Month,
ROUND(SUM(Quantity * Price), 2) AS Monthly_Revenue
FROM customer_clean_transactions GROUP BY Month ORDER BY Month;

SELECT * FROM Monthly_Revenue_Trend;
```

- Country_Wise_Churn_Risk_Summary: Joins transaction and churn prediction data to summarize churn risk by country, assisting in identifying geographic vulnerabilities.

```sql
CREATE VIEW Country_Wise_Churn_Risk_Summary AS
SELECT c.Country, COUNT(DISTINCT p.CustomerID) AS Total_Customers,
ROUND(AVG(p.Churn_Probability), 2) AS Avg_Risk FROM churn_predictions p
JOIN customer_clean_transactions c ON p.CustomerID = c.CustomerID
GROUP BY c.Country ORDER BY Avg_Risk DESC;

SELECT * FROM Country_Wise_Churn_Risk_Summary;
```

- Segment_Wise_Churn: Summarizes churn statistics (counts, average risk) across customer segments, useful for comparing churn behaviors across RFM clusters.

```sql
CREATE VIEW Segment_Wise_Churn AS
SELECT Segment, COUNT(*) AS Total_Customers,
SUM(CASE WHEN Predicted_Churn = 0 THEN 1 ELSE 0 END) AS Churned,
ROUND(AVG(Churn_Probability), 4) AS Avg_Risk FROM churn_predictions
GROUP BY Segment ORDER BY Avg_Risk DESC;

SELECT * FROM Segment_Wise_Churn;
```

Conclusion

This project successfully analyzed customer behavior, performed segmentation using RFM and KMeans, predicted churn using machine learning, and visualized business insights via a Power BI dashboard. These outcomes provide actionable intelligence for improving customer retention and reducing business loss due to churn.

Future enhancements may include:

- Hyperparameter tuning for XGBoost
- Feature enrichment from web/app logs
- Time-series forecasting of revenue and customer lifetime value

---

Appendix

GitHub: [GitHub]

Dashboard: [Dashboard]