Github link :- https://github.com/romitsoni/credit-card-churn-analysis

Credit Card Churn Prediction Report

Overview:->

This report explains how a machine learning (ML) model was used to predict which credit card customers might stop using the service. It looks at customer data to find signs of possible churn, helping businesses take steps to keep those customers and reduce losses.

Data Cleaning:->

Initial Data Loading and Inspection:->

Dataset: exl_credit_card_churn_data.csv containing customer information
Original Shape: The dataset's dimensions and column structure were first analyzed

Missing Values Analysis: Comprehensive check for null values across all columns

Imputed Missing Values with Median, Mod

Gender Column Cleaning:->

Problem: Inconsistent string representations and 'nan' string values

Solution:- Converted all values to lowercase and standardized gender representation, imputed the missing values with mod imputation

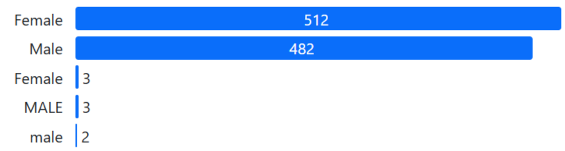
Before:->

Gender

Categorical

Imbalance

Distinct	6
Distinct (%)	0.6%
Missing	6
Missing (%)	0.6%
Memory size	61.3 KiB



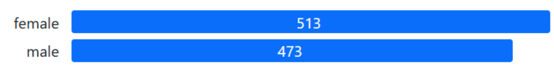
More details

After Data Cleaning:->

Gender

Categorical

Distinct	2
Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Memory size	59.9 KiB



More details

Age Column Cleaning:->

Problem: Negative, Missing age values found in the dataset

Solution: Identified rows with Age < 0, Replaced negative ages with np.nan
Imputed Missing values with median

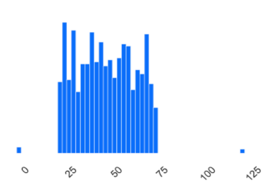
Before:->

Age

Age

Real number (ℝ)

Distinct	55	Minimum	-5
Distinct (%)	5.5%	Maximum	120
Missing	2	Zeros	0
Missing (%)	0.2%	Zeros (%)	0.0%
Infinite	0	Negative	3
Infinite (%)	0.0%	Negative (%)	0.3%
Mean	43.713294	Memory size	8.0 KiB



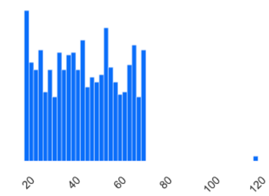
More details

After Data Cleaning:->

Age

Real number (ℝ)

Distinct	54	**Minimum**	18
Distinct (%)	5.4%	**Maximum**	120
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	43.862588	**Memory size**	7.9 KiB



More

:-> In the same way I handled any inconsistencies or negative values for the other columns

Churn Column (Target Variable) Cleaning:->

Problem: Mixed representations of churn status
Solution:

Mapped: '1', '1.0' → 1; '0', '0.0', '2', '2.0', 'maybe' → 0
Ensured binary classification format

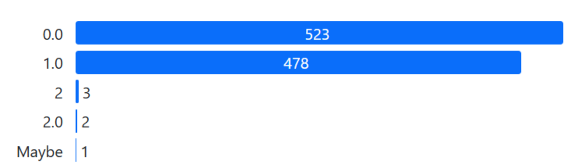
Before:->

Churn

Categorical

Imbalance

Distinct	5
Distinct (%)	0.5%
Missing	3
Missing (%)	0.3%
Memory size	59.3 KiB



More details

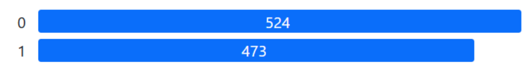
After Cleaning:->

Churn

Churn

Categorical

Distinct	2
Distinct (%)	0.2%
Missing	0
Missing (%)	0.0%
Memory size	48.8 KiB



More details

Duplicate Removal:->

Problem: There are duplicate rows present in the table

Solution:-Removed the duplicate rows

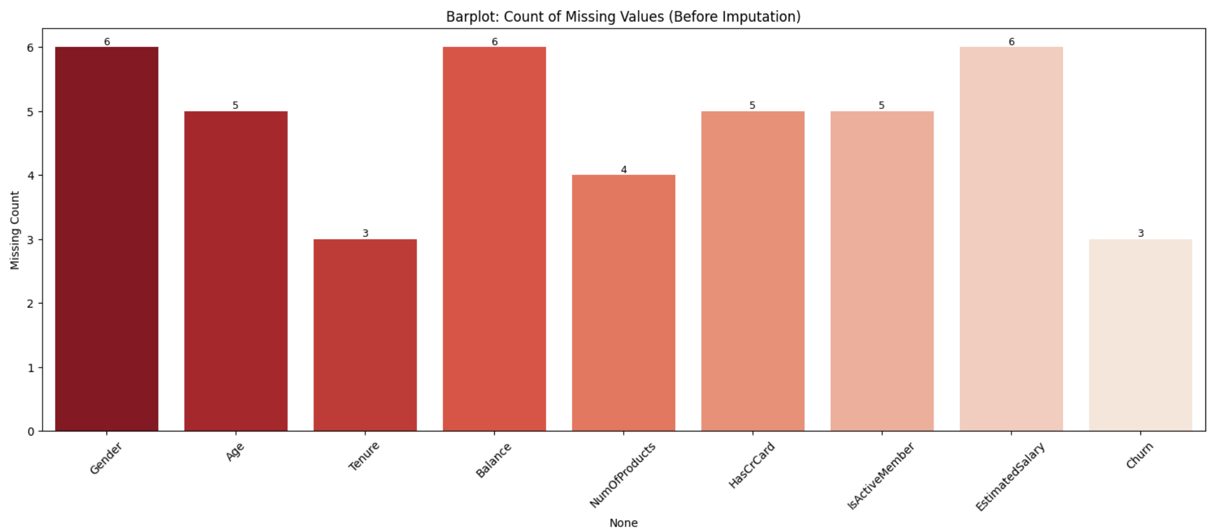
```
[Before] Total rows: 1010
Duplicate rows count: 10
Duplicate rows: 10
New shape after removing duplicates: (1000, 10)
```

Duplicate rows

Most frequently occurring

	CustomerID	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Churn	# duplicates
0	CUST0075	Male	64.0	5.0	0.00	3.0	1.0	0.0	52888.72	0.0	2
1	CUST0152	Male	57.0	3.0	0.00	3.0	1.0	1.0	37550.17	0.0	2
2	CUST0205	Male	58.0	10.0	88282.84	2.0	1.0	0.0	104538.24	1.0	2
3	CUST0211	Male	53.0	2.0	119946.99	2.0	0.0	1.0	148977.33	0.0	2
4	CUST0283	Male	48.0	8.0	0.00	1.0	1.0	0.0	46739.34	1.0	2
5	CUST0426	Male	39.0	4.0	0.00	1.0	0.0	1.0	98427.93	0.0	2
6	CUST0466	Female	46.0	3.0	0.00	1.0	1.0	0.0	75847.47	1.0	2
7	CUST0570	Female	34.0	7.0	187099.25	1.0	1.0	1.0	81839.42	0.0	2
8	CUST0629	Male	26.0	4.0	0.00	4.0	0.0	1.0	89569.80	1.0	2
9	CUST0704	Male	55.0	6.0	5765.31	2.0	1.0	0.0	42673.20	1.0	2

Bar Plot of Missing Values Before Imputation:->



Missing Value Imputation Strategy:->

Numerical Columns (Age, Tenure, Balance, NumOfProducts, EstimatedSalary): Imputed with median values

Categorical Columns (Gender, HasCrCard, IsActiveMember): Imputed with mode values

Target Variable: Rows with null Churn values were dropped entirely

```
 Imputing numerical columns with MEDIAN:
Imputed Age with median: 43.00
Imputed Tenure with median: 6.00
Imputed Balance with median: 6708.30
Imputed NumOfProducts with median: 2.00
Imputed EstimatedSalary with median: 83351.81

Imputing categorical columns with MODE:
Imputed Gender with mode: 'female'
Imputed HasCrCard with mode: '1'
Imputed IsActiveMember with mode: '0'
```

Churn Null Values Removal:->

```
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   CustomerID       1000 non-null   object
 1   Gender           1000 non-null   object
 2   Age              1000 non-null   Int64
 3   Tenure           1000 non-null   Int64
 4   Balance          1000 non-null   float64
 5   NumOfProducts    1000 non-null   Int64
 6   HasCrCard        1000 non-null   Int64
 7   IsActiveMember   1000 non-null   Int64
 8   EstimatedSalary  1000 non-null   float64
 9   Churn            997 non-null    Int64
```

Object Creation / Feature Engineering

```
After dropping null Churns:
CustomerID         0
Gender             0
Age                0
Tenure             0
Balance            0
NumOfProducts      0
HasCrCard          0
IsActiveMember     0
EstimatedSalary    0
Churn              0
dtype: int64
Final shape: (997, 10)
```

Outlier Detection and Handling:->

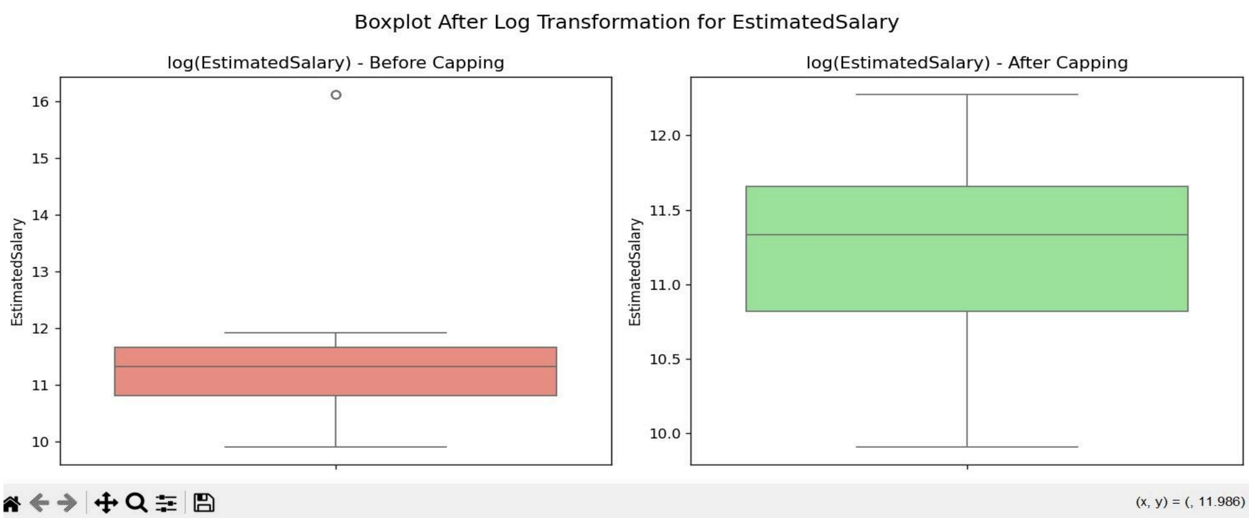
Clean extreme values that could affect model performance

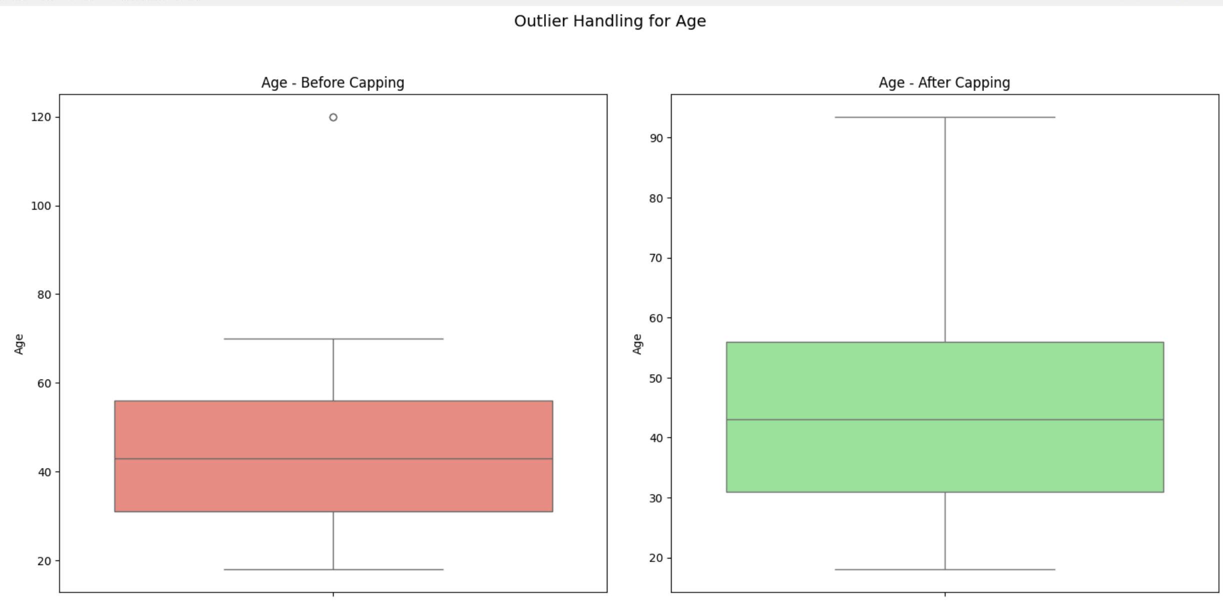
Method Used:-> IQR (Interquartile Range) Capping
Identified outliers in: Age, Balance, EstimatedSalary, NumOfProducts
Applied 1.5×IQR rule for boundary detection
Capped extreme values instead of removing rows

```
3.1 IQR METHOD WITH CAPPING:
-----------------------------------
Age: 0 outliers detected
  Lower bound: -9.00, Upper bound: 95.00
Balance: 2 outliers detected
  Lower bound: -165634.46, Upper bound: 276057.43
NumOfProducts: 0 outliers detected
  Lower bound: -3.50, Upper bound: 8.50
EstimatedSalary: 3 outliers detected
  Lower bound: -48733.08, Upper bound: 214375.45
```

Outlier Detection Visualizations:-

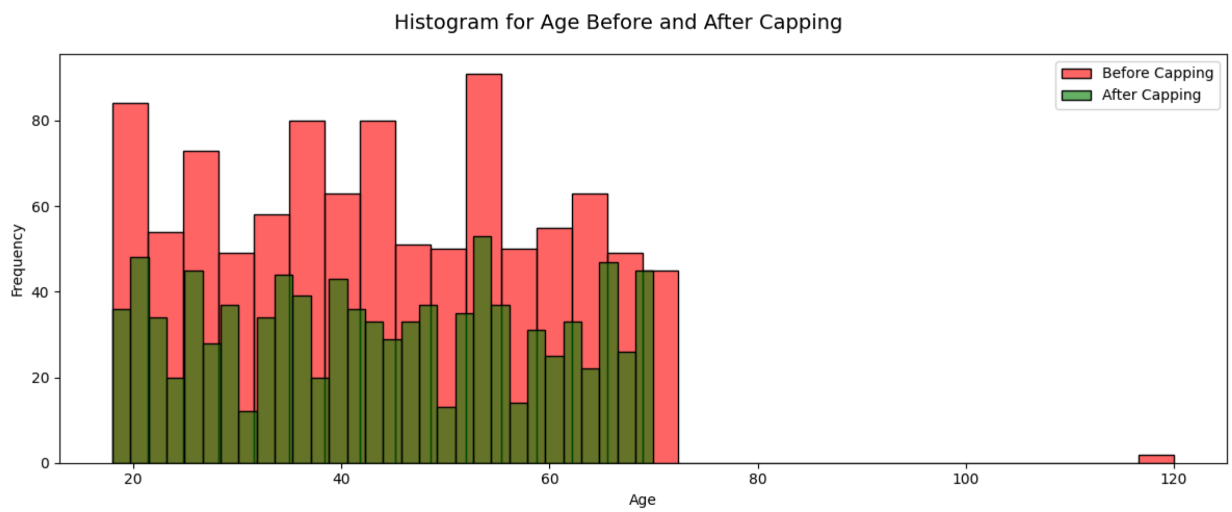
Box Plots: Before and after outlier capping comparison
Separate plots for Age, Balance, and EstimatedSalary
Log-transformed versions for skewed financial data

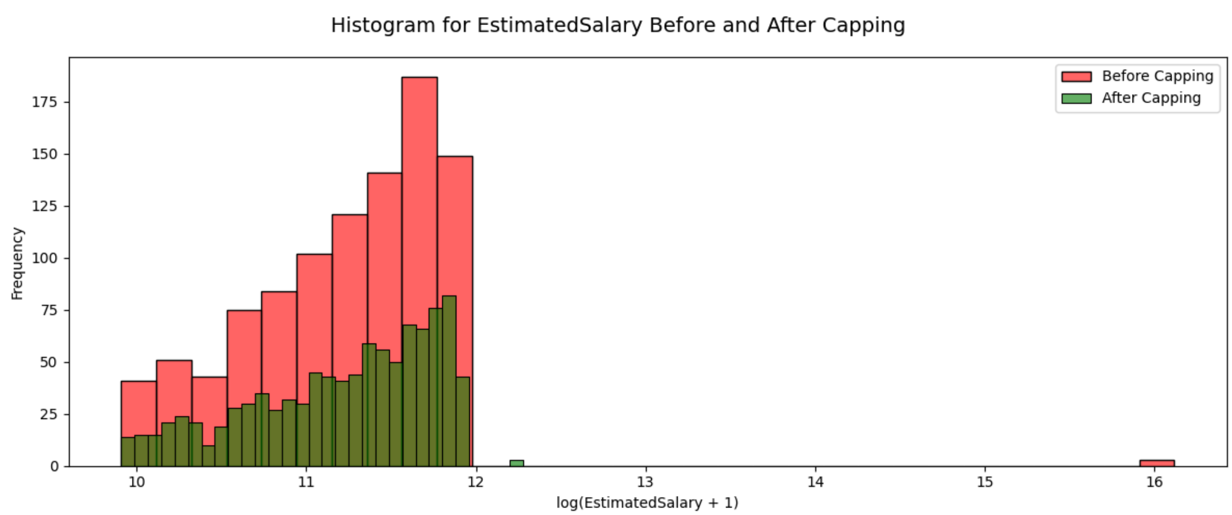




Histograms: - Compares age values before and after fixing outliers
Green bars show how extreme age values were limited

Helps make the data more balanced and realistic





Feature Engineering:->

To Create meaningful features to improve model performance

New Features Created:->

Financial Features:-

BalancePerProduct: Balance efficiency per product owned

SalaryPerProduct: Spending capacity per product

BalanceToSalaryRatio: Financial health indicator
LogBalance & LogSalary: Log-transformed financial features

Categorical Features:-
AgeBand: Age groups (18-29, 30-39, 40-49, 50-59, 60+)
TenureBand: Relationship length (New, Medium, Long)
ActivityLevel: Engagement level (Low, Medium, High)

Risk Features:-
HighValueCustomer: Top 25% by balance or salary
CustomerRiskScore: Composite risk indicator (0-1 scale)
HasZeroBalance: Flag for zero balance customers
IsSingleProduct: Flag for single product users

```
Data columns (total 22 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   CustomerID            986 non-null    object
 1   Gender                986 non-null    object
 2   Age                   986 non-null    int64
 3   Tenure                986 non-null    Int64
 4   Balance               986 non-null    float64
 5   NumOfProducts         986 non-null    int64
 6   HasCrCard             986 non-null    Int64
 7   IsActiveMember        986 non-null    Int64
 8   EstimatedSalary       986 non-null    float64
 9   Churn                 986 non-null    Int64
 10  BalancePerProduct     986 non-null    float64
 11  AgeBand               986 non-null    category
 12  TenureBand            986 non-null    category
 13  ActivityLevel         986 non-null    object
 14  SalaryPerProduct      986 non-null    float64
 15  BalanceToSalaryRatio  986 non-null    float64
 16  HighValueCustomer     986 non-null    int64
 17  CustomerRiskScore     986 non-null    float64
 18  HasZeroBalance        986 non-null    int64
 19  IsSingleProduct       986 non-null    int64
 20  LogBalance            986 non-null    float64
 21  LogSalary             986 non-null    float64
dtypes: Int64(4), category(2), float64(8), int64(5), object(3)
memory usage: 160.3+ KB
```

Exploratory Data Analysis (EDA):->

Key Findings:->

Overall Churn Rate: Calculated percentage of customers who churned vs. retained
Class Balance: Assessed whether the dataset has balanced or imbalanced target classes

Chart 1 – Overall Churn Distribution

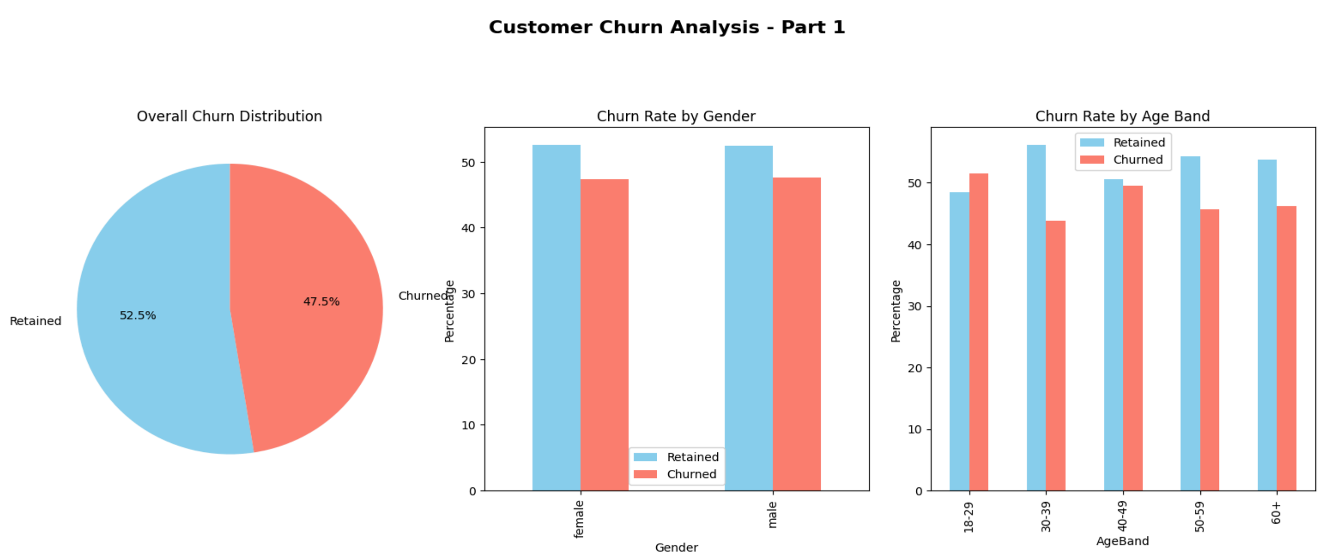
Around 47.5% of customers have churned
Churn rate is quite high, showing a need for better retention

Chart 2 – Churn Rate by Gender:->

Both males and females have similar churn patterns
Slightly more females are retained compared to males

Chart 3 – Churn Rate by Age Band:->

Younger (18–29) and older (60+) age groups churn more
Middle-aged groups (30–59) have higher retention
Age impacts churn, with certain age bands needing more focus



Behavioral Insights Visualization (Figure 2)

Chart 1 – Churn by Activity Level

Highly active customers are more likely to stay
Low activity customers churn the most
Shows that customer engagement plays a big role in retention

Chart 2 – Churn by Number of Products

Customers with only 1 product have the highest churn
Churn decreases as product count increases
Suggests that offering more products can improve loyalty

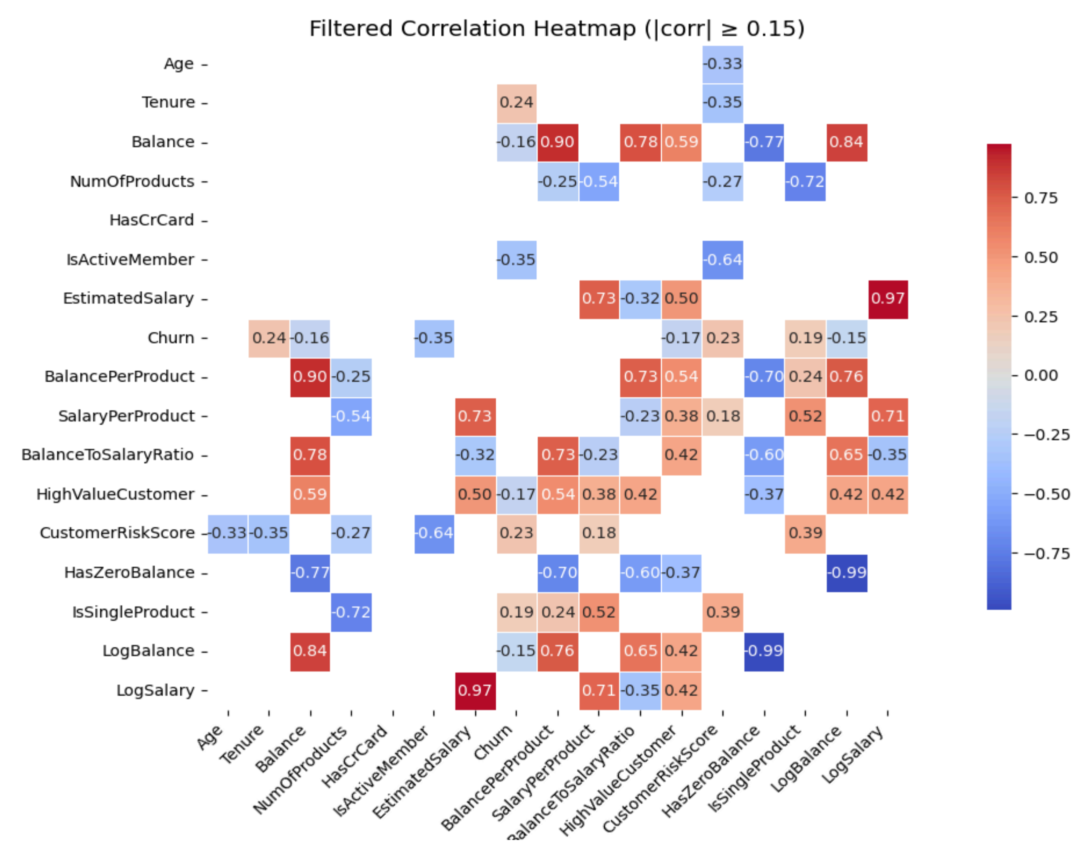
Chart 3 – Churn by Tenure Band

New customers (less than 2 years) churn more
Long-term customers (over 5 years) are more loyal
Indicates early months are crucial for customer retention

Correlation Analysis:->

Filtered Correlation Heatmap:-

Shows only strong relationships (correlation above 0.15)
Helps focus on patterns that matter most
Reveals how features like balance or salary relate to churn
Hides weak or unimportant connections to reduce confusion



Data Preparation:->

Steps Completed:

Applied one-hot encoding to categorical variables

Encoded: Gender, AgeBand, TenureBand, ActivityLevel

```
------------------------------
object-type columns (typically categorical):
CustomerID         object
Gender             object
ActivityLevel      object
dtype: object

category-type columns (explicitly categorized):
AgeBand        category
TenureBand     category
dtype: object

original number of features: 22
number of features after one-hot encoding: 27

one-hot encoded columns added:
 - Gender_male
 - AgeBand_30-39
 - AgeBand_40-49
 - AgeBand_50-59
 - AgeBand_60+
 - TenureBand_Medium (3-5 years)
 - TenureBand_Long (6+ years)
 - ActivityLevel_Low
 - ActivityLevel_Medium
```

Scaling:

Applied MinMaxScaler to all numerical features

Normalized values to 0-1 range for model compatibility

Train-Test Split

Split Ratio: 80% training, 20% testing

```
Train set: 788 samples
Test set: 198 samples
Train churn rate: 0.473
Test churn rate: 0.475
model training and evaluation
8.1 training models:
```

Model Training and Evaluation:->

Model Tested on two ML algorithms 1. Logistic Regression 2. Random Forest

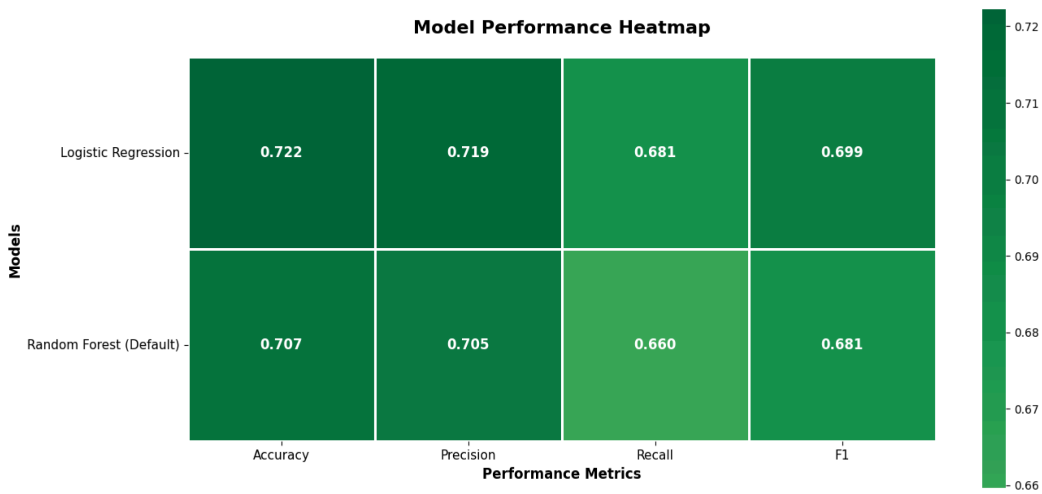
Logistic Regression:-
Accuracy: 72%
Churn Precision (class 1): 0.72
Churn Recall: 0.68
F1-Score (churn): 0.70
Performs slightly better at identifying churned customers
Balanced performance across both classes

Random Forest (Default):-
Accuracy: 71%
Churn Precision: 0.70
Churn Recall: 0.66
F1-Score (churn): 0.68
Slightly lower recall -> misses more churners

```
logistic regression - detailed metrics
--------------------------------------------------
classification report:
              precision    recall  f1-score   support

         0.0       0.72      0.76      0.74       104
         1.0       0.72      0.68      0.70        94

    accuracy                           0.72       198
   macro avg       0.72      0.72      0.72       198
weighted avg       0.72      0.72      0.72       198


random forest (default) - detailed metrics
--------------------------------------------------
classification report:
              precision    recall  f1-score   support

         0.0       0.71      0.75      0.73       104
         1.0       0.70      0.66      0.68        94

    accuracy                           0.71       198
   macro avg       0.71      0.70      0.71       198
weighted avg       0.71      0.71      0.71       198
```

Model Performance Heatmap:->

Chart Type: Heatmap for each model (Logistic Regression, Random Forest)
Data: True vs. Predicted classifications



Confusion Matrix Observations:->

Logistic Regression

Correctly predicted 79 retained customers (76%)

Correctly predicted 64 churned customers (68.1%)

Misclassified 25 retained as churned

Misclassified 30 churned as retained

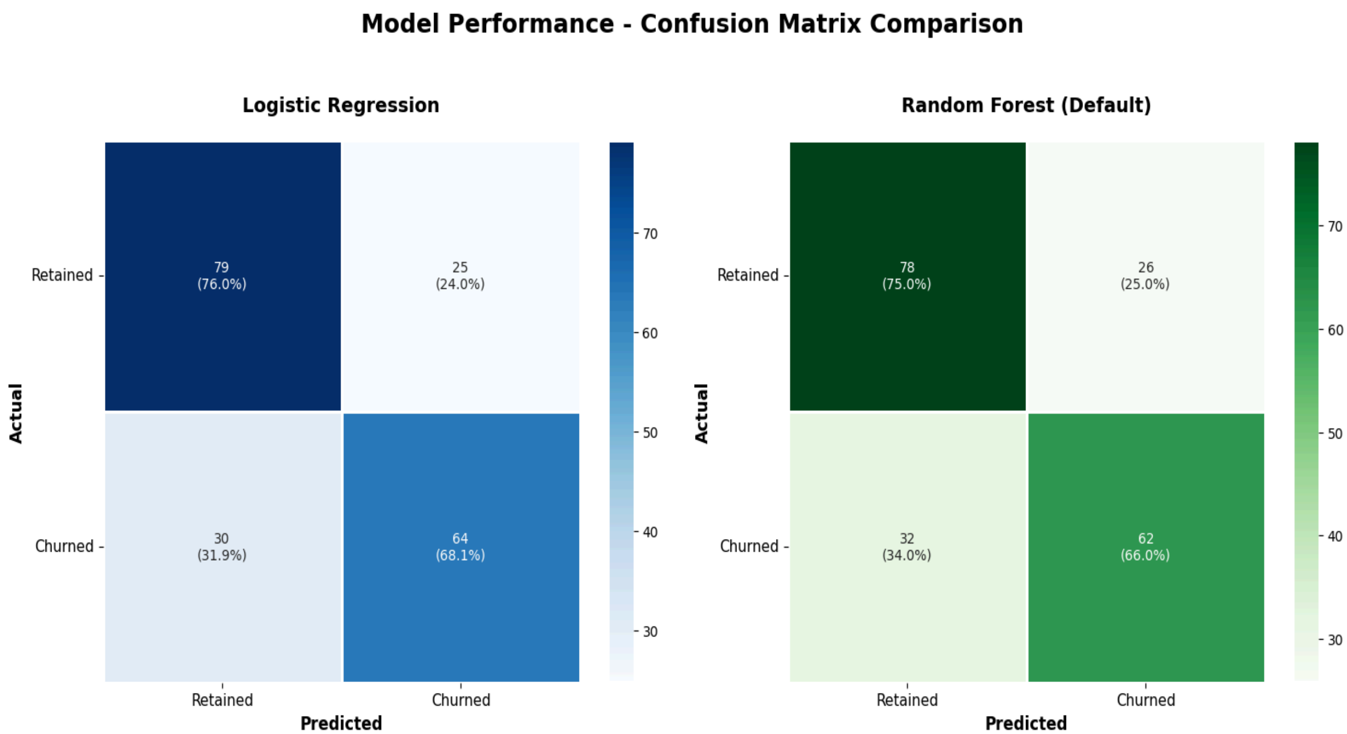
Random Forest:->

Correctly predicted 78 retained customers (75%)

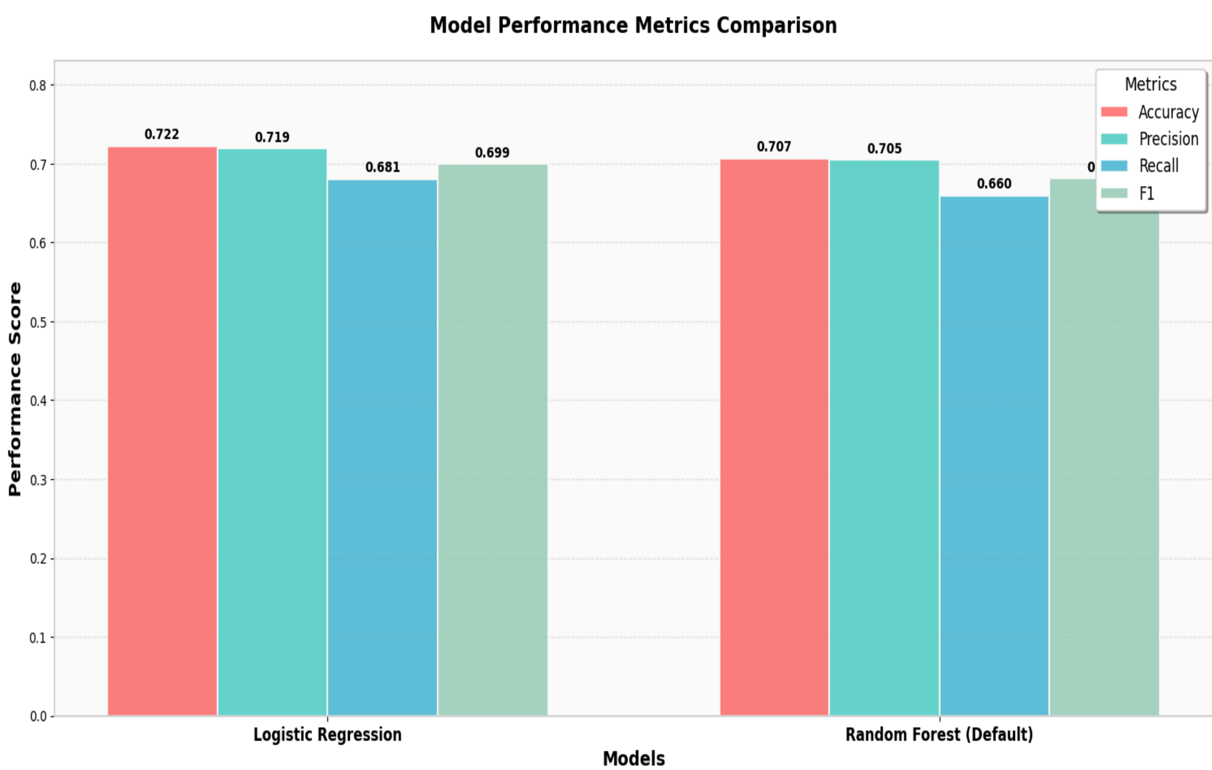
Correctly predicted 62 churned customers (66%)

Misclassified 26 retained as churned

Misclassified 32 churned as retained



Model Comparison Bar Chart:
Chart Type: Grouped bar chart
Metrics: Accuracy, Precision, Recall, F1-score



Business Insights:->

High-Risk Customer Segments:->

1. Single Product Users: Highest churn risk group
2. Young Customers (18-29): Need targeted retention
3. Low Activity Customers: Require engagement programs

```
10.2 BUSINESS INSIGHTS:
-------------------------
C:\Users\romit\Music\backup\credit-card-churn-analysis\scr
 behavior or observed=True to adopt the future default and
  age_churn_rates = df.groupby('AgeBand')['Churn'].mean().
Highest churn age group: 18-29 (51.5%)
Highest churn product count: 1 products (63.3%)
Highest churn activity level: Low (61.8%)

10.3 TOP RISK FACTORS:
```

Key Insights from Figure - Part 1:->

Churn Rate by Customer Value (Left Chart):->

Regular customers churn more (55%) than high-value ones (37.6%)

Since valuable customers stay longer, it makes sense to focus on retaining them

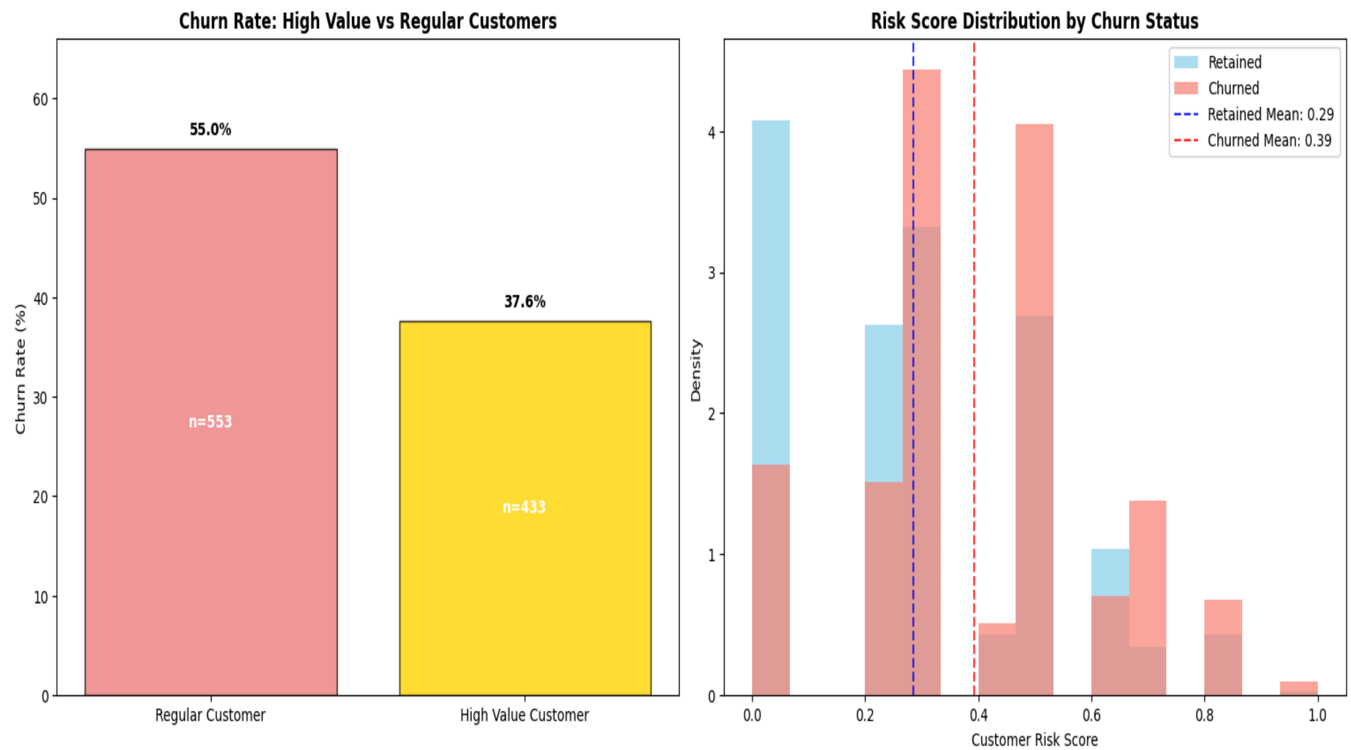
Customer Risk Score Distribution (Right Chart):->

Churned customers have a higher mean risk score (0.39)

Retained customers cluster around a lower mean risk score (0.29)

Good separation between churned and retained customers means this feature helps the model make better predictions

Financial Health & Risk Analysis - Part 1



Key Insights from Figure - Part 2:->

Left Chart: Churn Rate by Balance-to-Salary Ratio

Very Low ratio → Highest churn (53%) → Financially strained customers

High ratio → Lowest churn (29.5%) → Strong financial buffer

Customers with a good balance-to-salary ratio are more likely to stay

Right Chart: Churn by Product & Balance Pattern

Single Product + Zero Balance → Extreme churn (72.8%)
Adding balance or products significantly lowers churn
Multi Product + Has Balance → Lowest churn (35.5%)
Shows that having more products and a positive balance helps keep customers from leaving