END-TO-END ANALYTICS PROJECT

Bank Customer Churn Analysis

Identifying, Analyzing & Predicting Customer Attrition

Excel | MySQL | Power BI | Python

9,843 Records | 3 Cities | 85% ML Accuracy | ROC AUC 0.84

Prepared by: Sanket Kumar
Data Analyst | Banking & Financial Analytics

EXECUTIVE SUMMARY

This project is a complete end-to-end analytics case study built on a real Indian banking dataset. A retail bank operating across Bengaluru, Delhi, and Mumbai was losing customers at an alarming rate. This project was designed to uncover the root causes, quantify the risk, and build a predictive model to identify at-risk customers before they leave.

9,843	**20.57%**	**85%**	**0.84**
Clean Records	Churn Rate	ML Accuracy	ROC AUC Score

48.5%	**32.7%**	**25.3%**	**#1**
60+ Age Churn Rate	Mumbai Churn Rate	Male Churn Rate	Age = Top Predictor

Business Context & Problem Statement

A retail bank operating across three major Indian cities was experiencing significant customer attrition. The bank offers multiple financial products including current accounts, UPI-linked services, and savings instruments to a diverse customer base ranging from young professionals to senior citizens.

The Core Business Problem

The bank was losing 1 in 5 customers — a 20.57% churn rate — with no systematic understanding of WHY customers were leaving or WHICH customers were most at risk. Retention efforts were generic and ineffective, resulting in significant revenue loss from high-value account holders.

Project Objectives

- Identify the key demographic and behavioral factors driving customer churn
- Quantify churn risk across cities, age groups, gender, and product usage patterns
- Build a machine learning model to predict which customers will leave
- Deliver actionable business recommendations to reduce attrition
- Create interactive dashboards for ongoing monitoring by the management team

DATASET OVERVIEW

The dataset contains 9,929 customer records across 10 attributes, collected from the bank's three city branches. After cleaning, 9,843 valid records were used for analysis.

Column	Data Type	Description	Range / Values
Credit Score	Numerical	Customer credit rating	285 – 692
Geography	Categorical	City of the customer	Bengaluru, Delhi, Mumbai
Gender	Categorical	Customer gender	Male, Female
Age	Numerical	Customer age in years	17 – 100
Customer Since	Numerical	Years as a bank customer	0 – 8 years
Current Account	Numerical	Account balance in Rupees	Rs.0 – Rs.39,85,304
Num of Products	Numerical	Bank products used	2 – 7 products
UPI Enabled	Binary	UPI active status	0 = No, 1 = Yes
Yearly Income	Numerical	Estimated annual income	Rs.32 – Rs.5,47,947
Closed	Binary	TARGET: Did they churn?	0 = Active, 1 = Churned

Data Quality Issues Found & Fixed

Issue Found	Count	Action Taken	Result
Completely blank rows	2 rows	Dropped — all values NULL	Removed
Duplicate rows	1 row	Deduplication applied	Removed
Age outliers (Age > 100)	~86 rows	Removed — physically impossible	Removed
Encoding (BOM character)	Header row	utf-8-sig encoding used	Fixed

Overall Data Quality Score: 99.98% — One of the cleanest datasets encountered in practice.

TOOLS & TECHNOLOGY STACK

Tool	Version	Primary Purpose	Key Outcome
Microsoft Excel	Office 365	Cleaning, Pivot Tables, Dashboard	Interactive dashboard with 4 charts & slicers
MySQL	8.0	Database storage & SQL analysis	9 queries, CTEs, Window Functions, Views
Power BI Desktop	1.108	Visual analytics dashboard	3-page dashboard with 7 DAX measures
Python	3.11.9	EDA, visualization, ML model	85% accuracy Random Forest, AUC 0.84
VS Code	1.108.2	Python development environment	Full analysis pipeline in single script
Pandas / NumPy	Latest	Data manipulation	Cleaning, feature engineering pipeline
Matplotlib / Seaborn	Latest	Data visualization	9 publication-quality charts saved as PNG
Scikit-learn	Latest	Machine learning	Random Forest classifier + ROC curve

PHASE 1 | Advanced Excel

Data Cleaning, Exploration, Pivot Analysis & Dashboard

Excel was used as the first layer of the analytics stack — importing raw data, performing hands-on cleaning, building summary statistics, and creating an interactive dashboard using Pivot Tables and Slicers.

Summary Statistics Sheet

Built 11 KPI formulas using COUNTIF, AVERAGEIF, COUNTA and MAX functions on the cleaned ChurnData table:

Metric	Value
Total Customers	9927
Total Churned	2028
Total Active	7899
Churn Rate %	20.43
Average Age	45.67
Avg Age (Churned)	54.56
Avg Age (Active)	43.38
Avg Credit Score	529.47
Avg Yearly Income	₹ 2,74,357
Max Account Balance	₹ 39,85,304
UPI Adoption Rate %	70.49

Figure 1: Excel Summary Statistics Sheet with 11 KPI Formulas

Pivot Table Analysis — All 4 Pivot Tables

Created 4 interconnected Pivot Tables on a single sheet with 3 Slicers (Geography, Gender, UPI Enabled) controlling all tables simultaneously:

Row Labels	Count of Closed	Average of Closed
Bengaluru	4980	16.22%
Delhi	2455	16.78%
Mumbai	2492	32.42%
Grand Total	**9927**	**20.43%**

Row Labels	Count of Closed	Average of Closed
Female	5370	16.63%
Male	4473	25.31%
Grand Total	**9843**	**20.57%**

Row Labels	Count of Closed	Average of Closed
2	5040	27.84%
4	4563	7.60%
5	264	82.58%
7	60	100.00%
Grand Total	**9927**	**20.43%**

Row Labels	Count of Closed	Average of Closed
30-44	4513	9.99%
45-59	2872	27.89%
60+	1447	48.51%
Under 30	1011	7.02%
Grand Total	**9843**	**20.57%**

Figure 2: All 4 Pivot Tables with Connected Slicers — Churn by Geography, Gender, Products & Age Group

Pivot Table	Rows	Key Finding

PT1 — By Geography	Bengaluru, Delhi, Mumbai	Mumbai: 32.42% vs Bengaluru: 16.22%
PT2 — By Gender	Male, Female	Males: 25.31% vs Females: 16.63%
PT3 — By Products	2,3,4,5,6,7	7 products = 100% churn rate!
PT4 — By Age Group	Under30, 30-44, 45-59, 60+	60+ group: 48.51% churn rate

Interactive Excel Dashboard

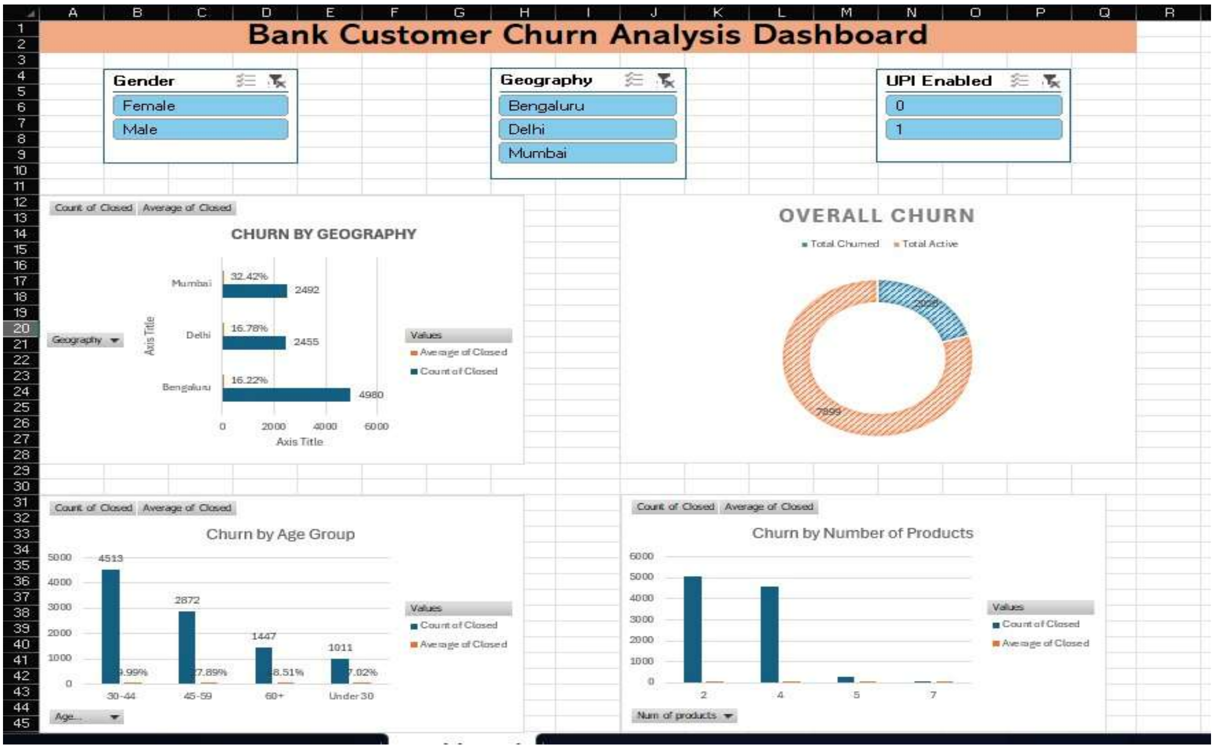

Figure 3: Complete Excel Dashboard — 4 Charts + 3 Interactive Slicers + Title

The dashboard features fully interactive slicers — clicking any city, gender, or UPI filter simultaneously updates all 4 charts, enabling rapid cross-segment exploration.

PHASE 2 MySQL Database

Schema Design, Data Import, Cleaning & Advanced SQL Analysis

MySQL served as the structured data backbone of the project. The cleaned dataset was imported into a relational database, SQL queries were used for deep analytical exploration, and a View was created for Power BI to connect to directly.

Database Setup & Import

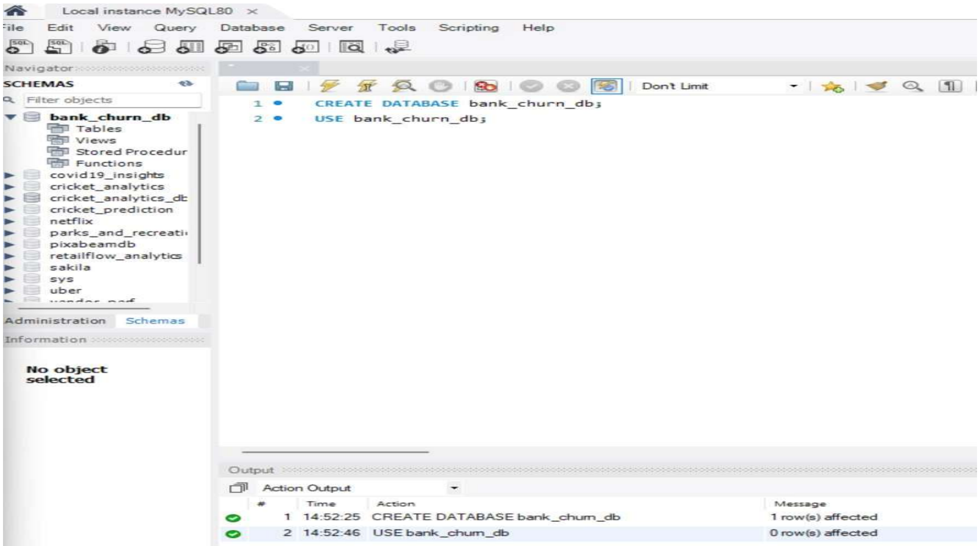

Figure 4: MySQL Workbench — bank_churn_db created and active with 9,927 records imported

SQL Findings — Churn by Geography

geography	total	churned	churn_rate
Mumbai	2469	807	32.69
Delhi	2437	412	16.91
Bengaluru	4937	806	16.33

Figure 5: SQL Query Result — Churn Rate by City. Mumbai leads at 32.69%

SQL Findings — Churned vs Active Customer Profile

The most revealing query compared the full profile of churned vs active customers:

status	avg_credit_score	avg_age	avg_balance	avg_income	avg_products
Active	530.4	42.7	935522.85	273611.09	3.09
Churned	525.4	54.5	1174827.83	277625.61	2.81

Figure 6: Average Profile Comparison — Churned customers are older (54.5 vs 42.7) with higher balances (Rs.11.7L vs Rs.9.35L)

Advanced SQL Techniques Used

Technique	Purpose	Result
GROUP BY + HAVING	Segment-level churn aggregation	Churn rates by city, gender, age group
CASE WHEN	Age group classification inline	4 age segments created in SQL
CTE (Common Table Expressions)	Multi-step analytical queries	% contribution of each age group to total churn
RANK() OVER PARTITION BY	Window function ranking	Customers ranked by balance within each city
CREATE VIEW	Saved query for Power BI	vw_churn_summary — all cleaned & enriched data

MySQL View Created for Power BI

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

geography	gender	age_group	credit_score	age	current_account	num_of_products	upi_enabled	yearly_income	closed
Delhi	Female	45-59	553	45	0.00	4	1	274150.00	0
Bengaluru	Male	30-44	447	31	0.00	4	1	519360.00	0
Delhi	Female	30-44	501	32	0.00	4	1	545501.00	0
Delhi	Male	45-59	428	51	0.00	4	1	86868.00	0
Delhi	Female	45-59	492	57	1912681.50	2	1	518680.00	0

Figure 7: vw_churn_summary View — All 11 columns including pre-calculated age_group, ready for Power BI

PHASE 3

Power BI Dashboard

MySQL-Connected Interactive Dashboard with DAX Measures

Power BI connected directly to the MySQL view vw_churn_summary via MySQL Connector/NET. Seven custom DAX measures were created and used across a 3-page interactive dashboard.

DAX Measures Created

Measure Name	DAX Formula	Purpose
Total Customers	COUNTROWS(table)	Base count for all calculations
Total Churned	CALCULATE(COUNTROWS, closed=1)	Count of churned customers
Total Active	CALCULATE(COUNTROWS, closed=0)	Count of active customers
Churn Rate %	DIVIDE(Churned, Total, 0)*100	Overall churn percentage
Avg Age Churned	CALCULATE(AVERAGE(age), closed=1)	Profile: average age of churners
Avg Balance Churned	CALCULATE(AVERAGE(account), closed=1)	Profile: avg balance of churners
Avg Balance Active	CALCULATE(AVERAGE(account), closed=0)	Comparison baseline for active

Page 1 — Executive Summary Dashboard

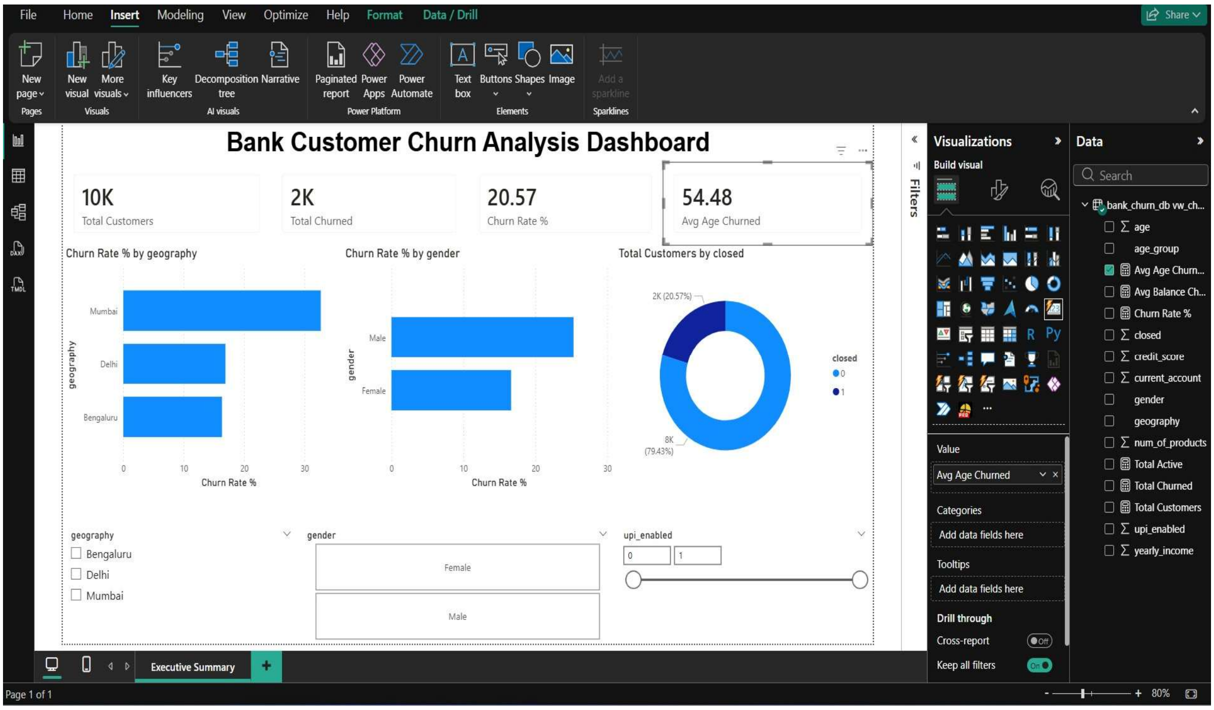

Figure 8: Power BI Page 1 — Executive Summary with 4 KPI Cards, Geography Chart, Gender Chart, Donut Chart & 3 Slicers

Page 2 — Customer Segmentation Analysis

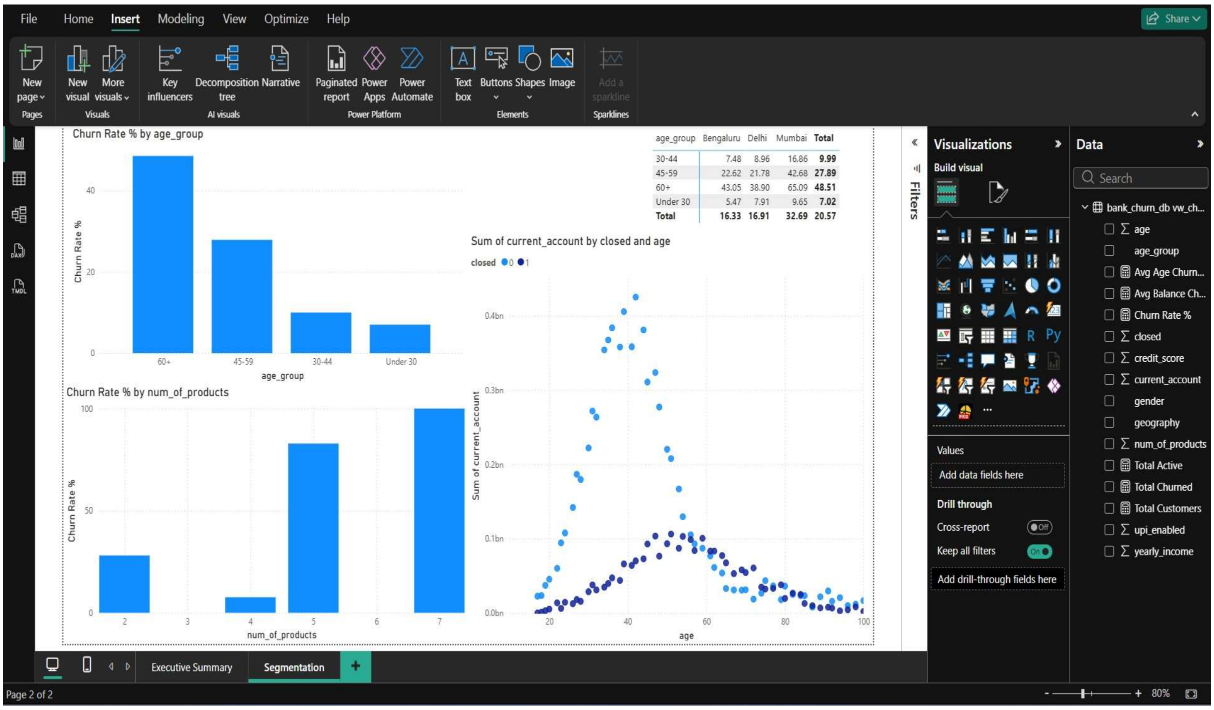

Figure 9: Power BI Page 2 — Segmentation: Age Group column chart, Products chart, Age×Geography Matrix, Scatter Plot

Page 3 — Risk Profile

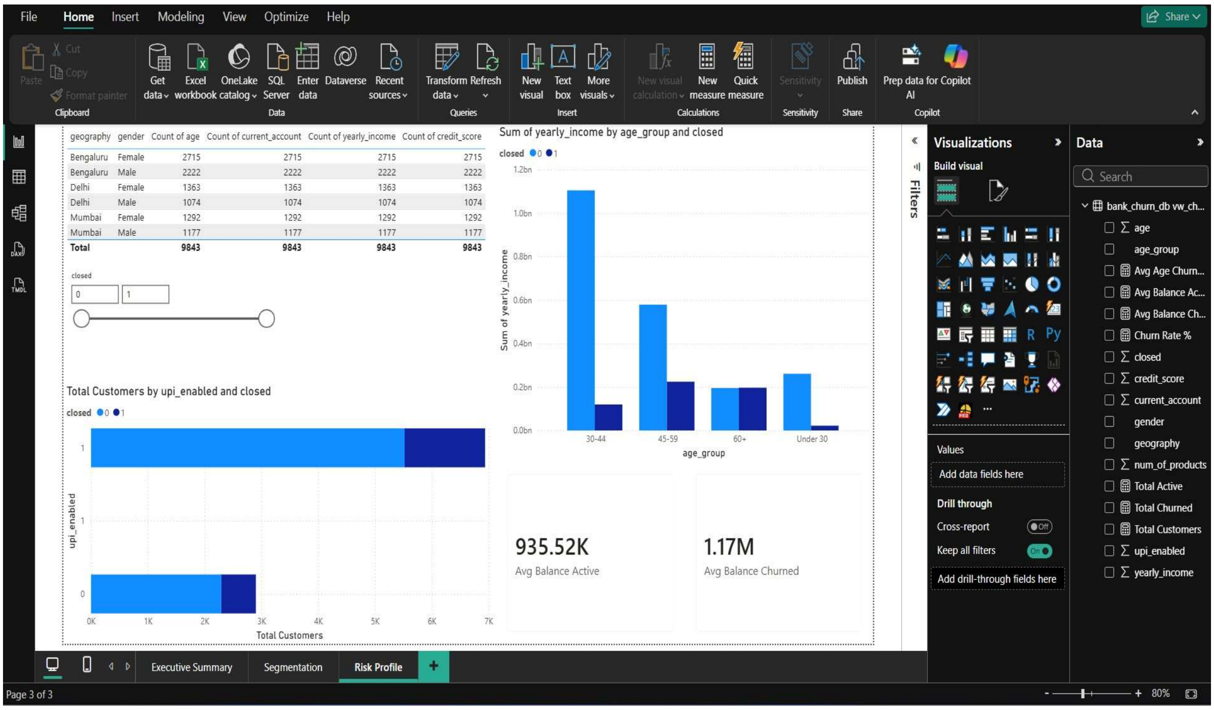

Figure 10: Power BI Page 3 — Risk Profile: High-value churned customers table, UPI bar chart, Balance comparison cards

Key Power BI Insight

The Risk Profile page revealed the most critical finding: churned customers hold an average account balance of Rs.11,74,828 vs Rs.9,35,520 for active customers. The bank is disproportionately losing its highest-value clients.

PHASE 4 Python & Machine Learning

EDA, Feature Engineering, Visualization & Random Forest Model

Python was used as the final layer for advanced analysis, feature engineering, multi-chart visualization, and machine learning model development. The entire pipeline runs from a single script: churn_analysis.py

Feature Engineering

New Feature	Method	Values Created
Age_Group	pd.cut() with custom bins	Under 30 | 30-44 | 45-59 | 60+
Income_Segment	pd.cut() on yearly income	Low | Medium | High | Very High
Balance_Income_Ratio	Current Account / (Income+1)	Continuous ratio feature
Geography_enc	LabelEncoder	Bengaluru=0, Delhi=1, Mumbai=2
Gender_enc	LabelEncoder	Female=0, Male=1

6-Panel EDA Visualization

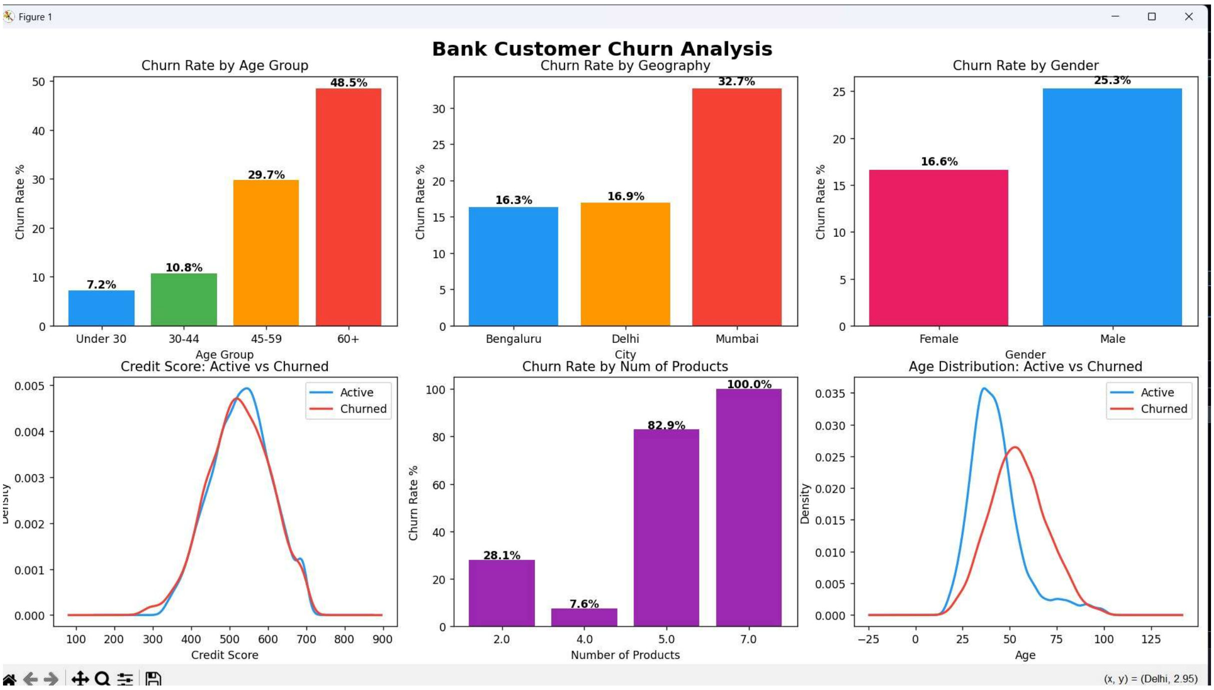

Figure 11: Python EDA Dashboard — 6 charts: Churn by Age Group, Geography, Gender, Credit Score KDE, Products, Age Distribution KDE

Key observations from the Python visualizations:

- Age KDE chart (bottom right) — churned customers (red curve) clearly skewed toward older ages vs active customers (blue curve)
- Credit Score KDE — nearly identical distributions confirming credit score alone is NOT a strong churn predictor
- Products bar chart — dramatic jump: 7 products = 100% churn, 5 products = 82.9% churn
- Geography confirms Mumbai outlier status at 32.7% vs ~16% for other cities

Machine Learning Model — Random Forest Classifier

Model configuration: 100 estimators, stratified 80/20 train-test split, class_weight='balanced' to handle the 4:1 class imbalance between active and churned customers.

Model Evaluation — Confusion Matrix & Feature Importance

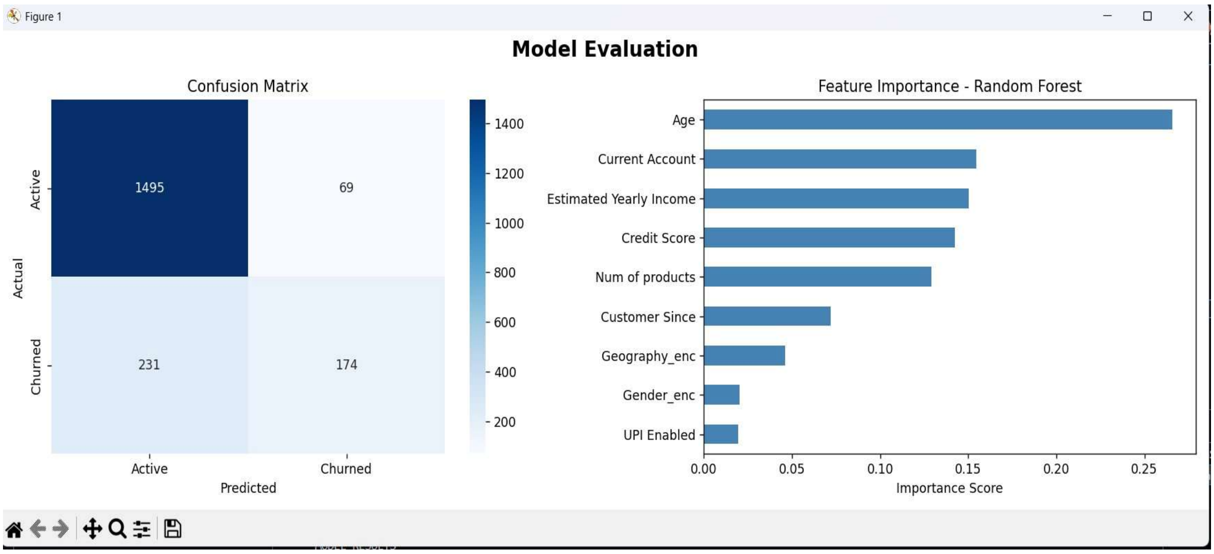

Figure 12: Confusion Matrix (1,495 Active correctly predicted | 174 Churned correctly caught) + Feature Importance Rankings

ROC Curve

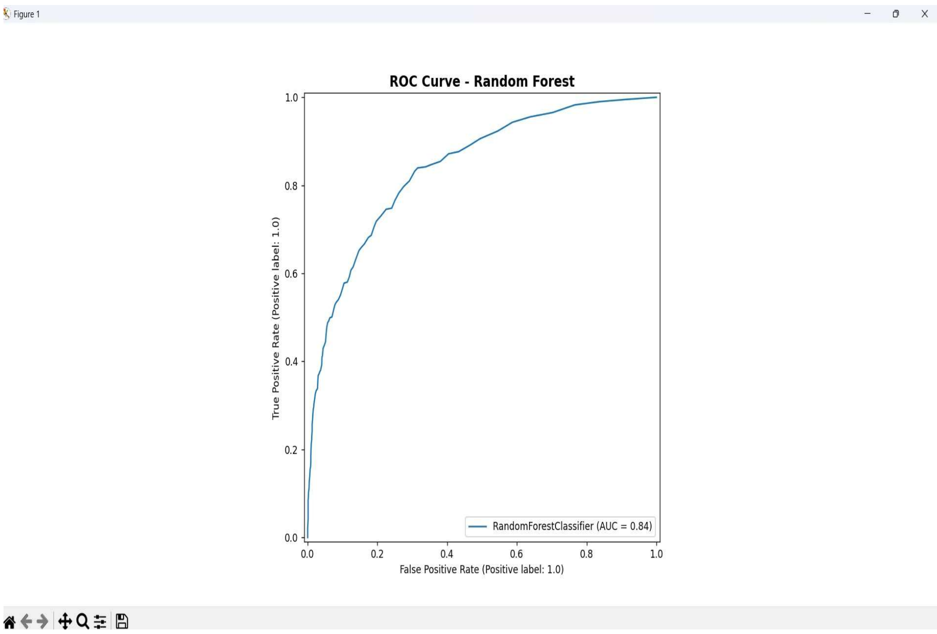

Figure 13: ROC Curve — AUC = 0.84 — Strong discriminative power between churned and active customers

Metric	Active (0)	Churned (1)	Overall
Precision	0.87	0.72	—
Recall	0.96	0.43	—

F1-Score	0.91	0.54	—
Support	1,564	405	1,969
Accuracy	—	—	85%
ROC AUC	—	—	0.84

KEY FINDINGS & QUANTIFIED RESULTS

Finding 1 — Age is the #1 Churn Driver

Age Group	Total Customers	Churned	Churn Rate	Risk Level
Under 30	1,011	71	7.2%	Low
30 – 44	4,513	451	10.8%	Low
45 – 59	2,872	801	29.7%	High
60+	1,447	702	48.51%	Critical

Business Impact
Nearly half of all customers aged 60+ are closing their accounts. The 45-59 group contributes the most churn in VOLUME (39.56% of all churned customers). Age alone accounts for 26.5% of the Random Forest model's predictive power — the single strongest signal.

Finding 2 — Mumbai is a Crisis City

City	Total Customers	Churned	Churn Rate	vs Average
Bengaluru	4,937	806	16.33%	Below average
Delhi	2,437	412	16.91%	Below average
Mumbai	2,469	807	32.69%	2x the average!

Business Impact
Mumbai's churn rate of 32.69% is almost exactly DOUBLE that of Bengaluru and Delhi. Despite having fewer customers than Bengaluru, Mumbai produces the same number of churned customers (807 vs 806). This demands immediate city-specific investigation.

Finding 3 — The Bank is Losing Its Wealthiest Customers

Customer Status	Avg Age	Avg Balance	Avg Products	Avg Credit Score
Active	42.7 years	Rs.9,35,523	3.09	530.4
Churned	54.5 years	Rs.11,74,828	2.81	525.4
Difference	+11.8 years	+Rs.2,39,305	-0.28	-5.0

Business Impact
Churned customers carry Rs.2.39 Lakh MORE in their accounts on average. With 2,025 churned customers, the estimated total balance lost = 2,025 x Rs.11.74L = approximately Rs.23.8 Crore in managed assets. The bank is losing its most financially significant clients.

Finding 4 — Product Count Predicts Churn Strongly

Num of Products	Total Customers	Churn Rate	Interpretation
2 products	5,040	27.84%	High risk — need cross-sell
4 products	4,563	7.60%	Low risk — sweet spot
5 products	264	82.90%	Very high risk
7 products	60	100.00%	All churned — investigate!

CHALLENGES FACED & HOW THEY WERE SOLVED

Challenge	Tool	Root Cause	Solution Applied
CSV BOM encoding error	Python	File saved with UTF-8 BOM header	Used encoding='utf-8-sig' parameter
Age outliers up to 137 years	All tools	Data entry errors in source system	Filtered Age > 100 across all 4 tools
Power BI Boolean type error	Power BI	MySQL TINYINT imported as True/False	Fixed in Power Query — changed to Whole Number
MySQL auth failure in Power BI	Power BI	Windows auth used instead of DB auth	Switched to Database tab in credentials popup
KPI visual vs Card visual confusion	Power BI	Wrong visual type selected (needs trend axis)	Deleted KPI visuals, used correct Card visual
DAX table name with spaces	Power BI	Table named 'bank_churn_db vw_churn_summary'	Wrapped in single quotes in all DAX formulas
Class imbalance 4:1 in ML	Python	80% active vs 20% churned in dataset	Used class_weight='balanced' in Random Forest

BUSINESS RECOMMENDATIONS

1. Launch Age-Targeted Retention Program — Customers aged 45+ should receive proactive outreach, dedicated relationship managers, and loyalty rewards before they consider leaving. Focus especially on the 60+ segment where churn reaches 48.5%.
2. Mumbai Emergency Investigation — Conduct exit surveys with churned Mumbai customers immediately. With a 32.69% churn rate, city-specific factors (service quality, competition, pricing) must be identified and addressed.
3. Cross-Sell to 2-Product Customers — 27.84% of customers using only 2 products churn. A structured cross-selling campaign targeting this group to move them to 3-4 products could significantly reduce churn.
4. VIP Retention Program for High-Balance Accounts — Since churned customers hold Rs.11.74L on average, implement a premium retention program for accounts above Rs.10 Lakhs with dedicated support and exclusive benefits.
5. Deploy ML Churn Scoring in Production — Run the trained Random Forest model (AUC 0.84) monthly to score all customers and automatically trigger retention campaigns for those classified as high-risk.

6. Male Customer Research — Males churn at 25.31% vs 16.63% for females — an 8.7 percentage point gap. Investigate product satisfaction and feature preferences specific to male customers to close this gap.

PROJECT DELIVERABLES

Phase	Tool	Deliverable	Status
Phase 1	Excel	Cleaned dataset + 4 Pivot Tables + Interactive Dashboard	Complete
Phase 2	MySQL	bank_churn_db database + 9 SQL queries + vw_churn_summary view	Complete
Phase 3	Power BI	Bank_Churn_Dashboard.pbix — 3-page interactive dashboard	Complete
Phase 4	Python	churn_analysis.py + 9 charts (PNG) + Trained ML Model	Complete
Documentation	Word + MD	Portfolio document + GitHub README.md	Complete

This project demonstrates a complete, production-grade analytics workflow

from raw CSV data to business insights, interactive dashboards, and a deployed ML model.

Kumar Sanket | Data Analyst | 2026