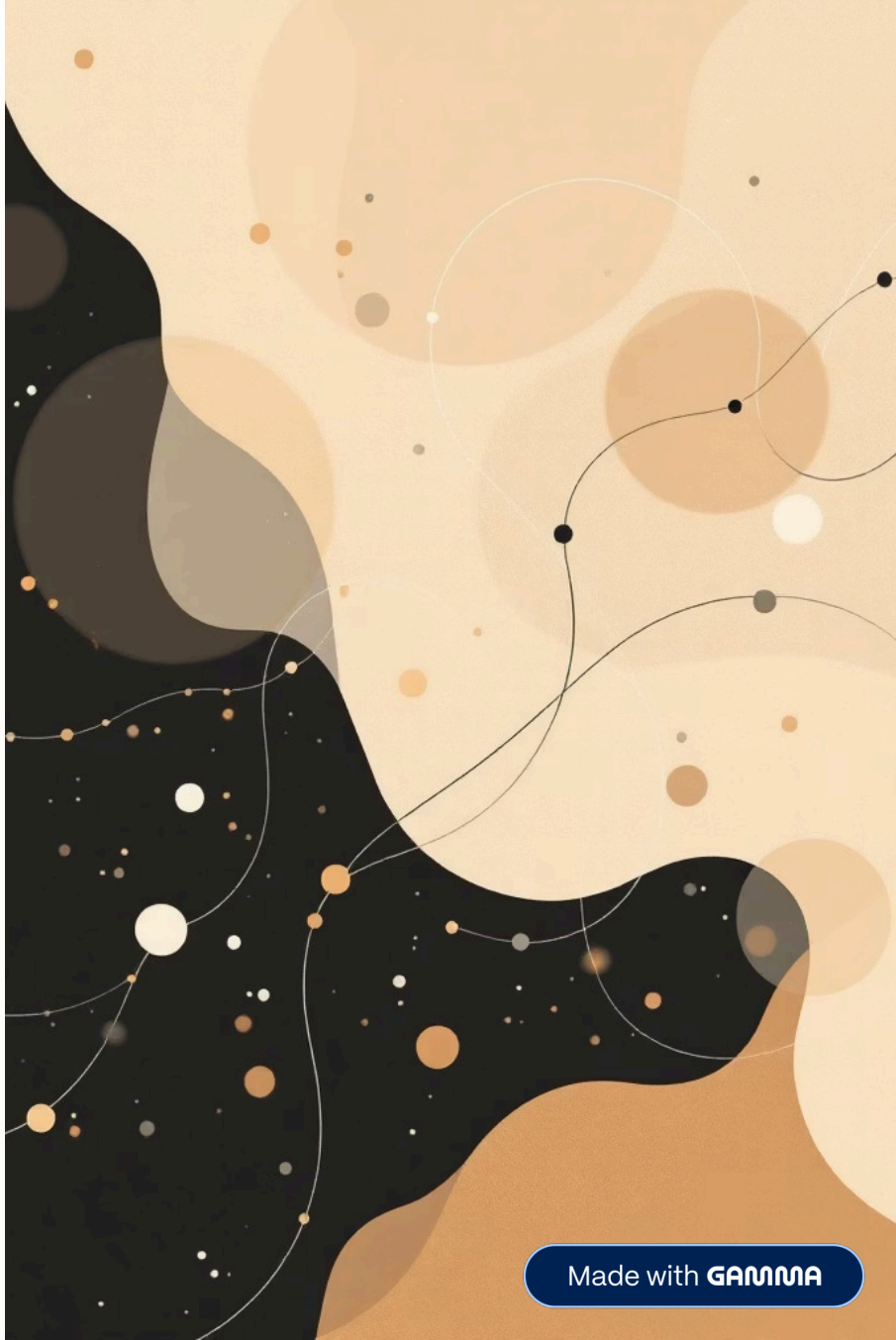
Customer Churn Prediction

An end-to-end machine learning pipeline to identify at-risk customers and enable proactive retention strategies.



CHAPTER 1 OVERVIEW

Project Overview

This project predicts customer churn using machine learning — identifying customers likely to leave so businesses can take **proactive retention measures**.

The pipeline covers data preprocessing, feature engineering, model comparison, evaluation, and prediction on new customer data.

CHAPTER 2 DATA

Dataset Description

The dataset contains **10,000 customer records** with the following features:

Customer Profile

credit_score, country, gender, age, tenure

Financial Data

balance, estimated_salary, products_number, credit_card

Target Variable: churn

0 → Customer did not churn

1 → Customer churned

CHAPTER 3 PREPROCESSING

Data Preprocessing

Remove

Dropped unnecessary columns like customer_id

Impute

Median for numeric, most frequent for categorical

Scale

Standard scaling for numeric features

Encode

One-hot encoding for categorical features

CHAPTER 3 FEATURE ENGINEERING

Engineered Features

Additional features were created to improve model interpretability and performance.

balance_per_product

Balance divided by number of products held

salary_balance_ratio

Ratio of salary to account balance

age_group

Categorical bucketing of customer age

tenure_bucket

Grouped tenure into meaningful ranges

high_balance

Flag for customers with elevated balances

CHAPTER 4 MODELING

Model Training & Comparison

Five classification models were compared using **Stratified K-Fold Cross Validation (5 folds)** with **ROC-AUC** as the evaluation metric.

The best model was selected by mean ROC-AUC score and trained on full training data.

01

Logistic Regression

02

Random Forest

03

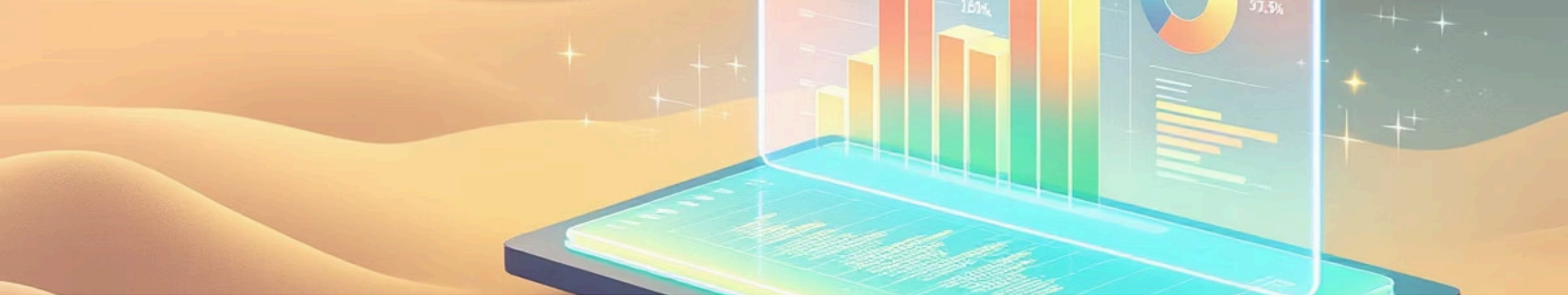
Gradient Boosting

04

AdaBoost

05

Support Vector Classifier



CHAPTER 5 EVALUATION

Model Evaluation Metrics

Accuracy & Precision

Overall correctness and positive prediction quality

Recall & F1-Score

Ability to catch churners and balanced performance

ROC-AUC & Confusion Matrix

Discrimination power and detailed classification report

Feature importance was extracted from tree-based models and visualized via bar plot to identify the most influential churn variables.

CHAPTER 6 PREDICTION

New Customer Prediction

The project includes real-world applicability:

1. Accept new customer data
2. Apply same feature engineering
3. Generate churn prediction & probability

CHAPTER 7 LEARNINGS & FUTURE

Key Learnings & Future Improvements

Key Learnings

- Proper preprocessing pipelines
- Preventing data leakage with sklearn Pipeline
- Handling imbalanced data with StratifiedKFold
- Using ROC-AUC for churn problems
- Interpreting results via feature importance

Future Improvements

- Hyperparameter tuning (GridSearchCV / RandomizedSearchCV)
- Threshold optimization for business tradeoffs
- Deployment via Streamlit or Flask
- Integration with real-time data sources

Conclusion

> This project demonstrates an end-to-end ML pipeline for customer churn prediction — from data preparation and feature engineering to model comparison, evaluation, and real-world prediction.

Technologies

Python, Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn

Best Practices

Industry-standard workflow suitable for practical business implementation