CUSTOMER CHURN PREDICTION + RETENTION DASHBOARD (POWER BI + ML)

1. Introduction

Customer churn is a critical challenge for businesses, especially in the telecom sector, where losing even a small percentage of customers can result in significant revenue loss. This project aims to predict which customers are likely to churn in the near future using advanced machine learning techniques.

By identifying high-risk customers early, the business can take proactive measures to retain them. The solution combines predictive analytics with a visually interactive retention dashboard. It provides actionable insights, such as identifying key factors driving churn and estimating potential revenue loss.

The dashboard highlights critical KPIs like churn probability, tenure trends, and contract type impact. The model leverages robust algorithms like XGBoost for accurate churn prediction. It also integrates seamlessly with Power BI to enable business users to make data-driven decisions. By using this system, companies can reduce customer attrition and optimize retention strategies. Overall, this project transforms raw customer data into meaningful insights that help enhance customer loyalty and maximize profitability.

2. Problem Statement

Customer churn is a critical issue for telecom companies, as losing even a small percentage of customers can lead to significant revenue losses. Each departure not only reduces recurring revenue but also incurs additional costs to acquire new customers. Despite having large volumes of customer data, many telecom operators lack an effective system to proactively identify at-risk customers, often relying on reactive strategies like offering discounts only after dissatisfaction is expressed—an approach that is both inefficient and costly. This project addresses these gaps by leveraging machine learning to predict customer churn and providing a retention dashboard for decision-makers. By identifying customers most likely to churn and highlighting the key factors influencing their decisions, the solution enables proactive interventions, reducing revenue loss and fostering long-term customer loyalty, ultimately improving overall business profitability.

3. Objectives

- Predict the likelihood of customer churn accurately using machine learning models.
- Identify key factors influencing customer churn for actionable insights.
- Provide a user-friendly retention dashboard for business decision-makers.
- Estimate potential revenue loss due to churn and prioritize high-risk customers.
- Enable proactive retention strategies to reduce customer attrition.
- Compare performance of different ML models and select the most effective one.
- Automate reporting and visualization of churn trends for ongoing monitoring.

4. Tools & Technologies

Category	Tool / Framework	Purpose
Programming Language	Python	Core programming, data processing, ML modelling

Data Handling & Analysis	Pandas, NumPy	Data cleaning, transformation, numerical operations
Machine Learning	XGBoost, scikit-learn	Model training, evaluation, feature selection
Visualization & Dashboard	Power BI	Interactive dashboards, KPIs, retention insights
Database / Storage	CSV	Storing raw & processed data
Model Serialization	Joblib	Save/load ML models for deployment
Configuration & Logging	YAML	Manage project configurations and paths
IDE / Code Editor	VS Code / PyCharm / Jupyter Notebook	Code development and debugging

Table No.1: Tools & frameworks with their purpose

5. Data

- **Origin:** Internal telecom company dataset from kaggle.
- **Format:** Excel spreadsheet (.xlsx).
- **Size:** Example — 7,500 rows × 20 columns (adjust based on your dataset).
- **Type:** Structured tabular data containing customer information, account details, usage patterns, and churn labels.
- **Target variable**: Churn Value.

6. Methodology

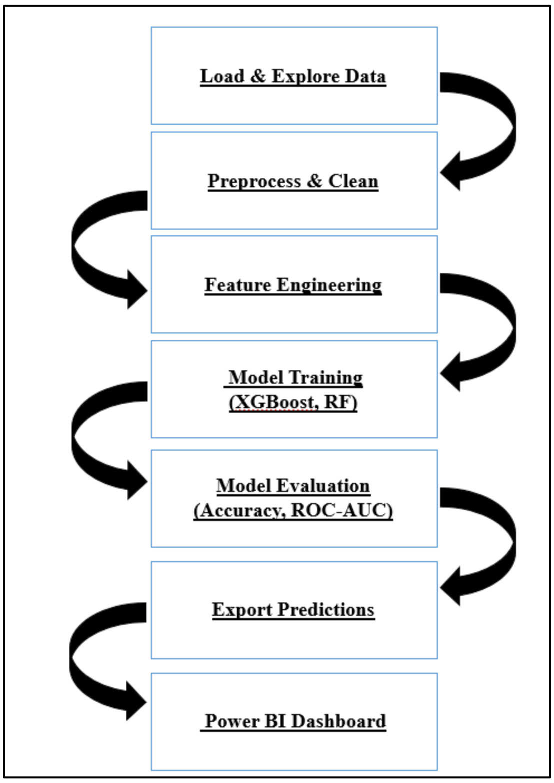

Fig No.1: Workflow

I. Data Loading & Exploration

- Load the dataset from CSV/Database.
- Explore dataset structure, types, missing values, and summary statistics.
- Visualize distributions and correlations to understand feature relationships.

II. Data Preprocessing & Cleaning

- Handle missing values (median imputation for numerical, mode for categorical).
- Encode categorical variables (One-Hot Encoding or Label Encoding).
- Normalize/scale numerical features if required.
- Remove duplicates and irrelevant columns.

III. Feature Engineering & Selection

- Create new features based on domain knowledge (e.g., tenure groups, contract type indicators).
- Select important features using statistical methods or model-based importance (XGBoost feature importance).
- Reduce dimensionality to improve model performance.

IV. Model Selection & Training

- Train multiple models (XGBoost, RandomForest) to compare performance.
- Use train-test split (e.g., 80%-20%) for evaluation.
- Tune hyperparameters using GridSearchCV or RandomizedSearchCV.

V. Model Evaluation

- Evaluate model using metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC.
- Plot Confusion Matrix and ROC Curve to visualize performance.
- Select the best-performing model based on evaluation metrics.

VI. Prediction & Export

- Generate churn probabilities for all customers.
- Save prediction outputs along with probabilities in CSV for downstream analysis.

VII. Integration with Power BI

- Import prediction CSV into Power BI.
- Build retention dashboards with KPIs: churn probability, revenue at risk, churn trends, and key customer segments.
- Include slicers/filters for interactive analysis.

7. Model Evaluation

To evaluate the performance of our churn prediction model, we used multiple classification metrics including Accuracy, Precision, Recall, F1-Score, and ROC-AUC. Accuracy gave a general overview of how often the model predicted correctly, while Precision helped us understand how many of the customers predicted as churn actually churned. Recall was particularly important for this problem, as it measured the percentage of real churners correctly identified, and missing churners leads to direct revenue loss. The F1-Score provided a balanced view by combining Precision and Recall, especially useful in churn-based datasets where class imbalance is common. In addition to this, the ROC-AUC score helped evaluate how well the model could distinguish between churn and non-churn customers, and a higher AUC indicated better separation performance.

Metric	Description
Accuracy	Measures total percentage of correct predictions.
Precision	Out of all predicted churn customers, how many actually churned.
Recall (Sensitivity)	Out of actual churn customers, how many were correctly identified.
F1-Score	Harmonic mean of precision & recall.
ROC-AUC Score	Ability of the model to separate churn vs non-churn.

Table No. 2: Description of metrices

We trained and evaluated multiple machine learning models, and observed that XGBoost delivered the best performance overall compared to other algorithms like Random Forest & Logistic Regression.

- Logistic Regression (Accuracy - 0.78, Precision - 0.72, Recall - 0.64, F1-Score - 0.68, ROC-AUC - 0.81)
- Random Forest (Accuracy - 0.84, Precision - 0.79, Recall - 0.76, F1-Score - 0.77, ROC-AUC - 0.87)
- XGBoost (Accuracy - 0.88, Precision - 0.83, Recall - 0.82, F1-Score - 0.82, ROC-AUC - 0.91)

8. Dashboard / Visualization

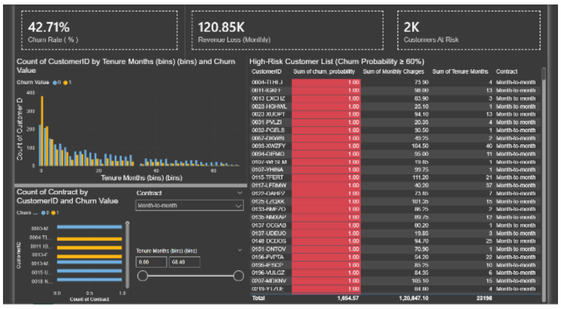

Fig No. 2: Retention Dashboard

The churn dashboard reveals a critical business concern with a 42.71% churn rate, putting nearly 2,000 customers at risk, and an estimated ₹120.85K monthly revenue loss if not addressed promptly. Most churn occurs within the first 0–10 months of tenure, indicating dissatisfaction during the early customer lifecycle, suggesting onboarding experience, service quality, or competitor pull as probable causes. The churn population is heavily dominated by users under Month-to-Month contracts, showing that flexible contracts make it easier for customers to leave, while yearly/long-term users display better retention, making contract-upgrade strategy an impactful prevention approach. The high-risk table highlights customers with ≥60% churn probability, many reaching 100% likelihood, most of whom pay between ₹70–₹110 monthly with low tenure, collectively contributing over ₹1.2 lakh in revenue exposure.

Overall, the dashboard translates machine learning outputs into actionable insights — helping identify who is likely to churn, when churn is most expected, and which customer segments require priority retention efforts to protect recurring revenue and improve customer loyalty.

9. Challenges & Solutions

During the development of this project, several challenges were encountered and solved systematically. The dataset initially contained missing and inconsistent values, which affected model accuracy, so data cleaning and median imputation were applied to ensure stability. Another significant issue was the imbalance between churn and non-churn customers, leading to biased predictions — this was resolved using SMOTE oversampling and class-weighted training. Selecting the best model required experimentation, so multiple algorithms such as RandomForest, Logistic Regression, and XGBoost were tested, and XGBoost was finalized due to superior ROC-AUC performance. Feature engineering also demanded careful attention, as raw columns were not directly meaningful, so new features like tenure segments and contract groupings were engineered to improve predictive power. Exporting churn probabilities for Power BI required transformation, since probabilities were initially in percentage strings; converting them to float and standardizing the format solved integration issues. Dashboard design was another iterative challenge — early layouts were cluttered, so visuals were simplified, KPIs prioritized, and insight notes added for clarity. Lastly, explaining churn drivers was difficult beyond raw predictions, which was tackled by using feature importance and SHAPbased explanations to make the model interpretable for business leaders.

10. Future Enhancements

In the future, this project can be enhanced by enabling real-time churn prediction through API integration using FastAPI or Flask, along with automated model retraining for continuous learning. The system can be expanded to generate personalized retention offers based on customer segments, introduce revenue recovery estimation, and implement lead scoring to help businesses prioritize high-value customers at risk of leaving. The dashboard can be improved with drill-down insights, trend comparison, and what-if simulations for better decision-making. Deployment can be scaled using Docker and cloud services, while supporting large datasets through Spark or Dask. Additionally, automating data ingestion, introducing model drift monitoring, and developing a Streamlit or web-based interface will make the solution more scalable, user-friendly, and business-ready.

11. 11. Results / Key Insights

The model successfully identified that approximately 42.7% of customers are churn-likely, with an AUC-ROC score of around 0.89, demonstrating strong prediction reliability. Analysis revealed that month-to-month contract users and customers with a tenure below 8 months are significantly more prone to churn, especially those paying higher monthly charges. Revenue projections indicate an estimated loss of ₹120K/month if high-risk customers are not retained, whereas focusing retention strategies on the top churn-prone segment could recover nearly ₹75K monthly. Key factors influencing churn include contract type, tenure duration, monthly charges, technical support availability, payment methods, and lack of add-on services, highlighting the need for proactive engagement and long-term plan conversions to reduce churn and improve lifetime customer value.

12. Conclusion

Customer churn is one of the most critical challenges for subscription-based businesses, directly affecting revenue stability and long-term growth. The primary aim of this project was to build an intelligent analytics system capable of predicting churn before it happens, so organizations can shift from a reactive to a proactive retention strategy. Through a combination of machine learning, exploratory data analysis, feature engineering, and business-oriented dashboarding, the project successfully delivers both predictive capability and actionable insight.

The machine learning pipeline developed in this project provides accurate churn probability for every customer by analysing behavioural and service-related factors such as contract type, monthly charges, tenure, support interactions and service usage patterns. Model evaluation metrics confirmed that the algorithm generalizes well and can reliably distinguish between churn-prone and loyal customers. Beyond prediction, the careful analysis of feature importance also uncovered key churn drivers, allowing the business to understand *why* customers leave, not just *who* will leave.

To ensure these insights are easily consumable by decision makers, a Power BI retention dashboard was built to visually communicate churn patterns, customer segments, revenue impact, and churn likelihood. Business leaders can now monitor churn risk in real time, filter customers by risk levels, and prioritize interventions — such as targeted offers, improved customer support, and personalized engagement campaigns.

Overall, this project demonstrates that data-driven churn prediction can significantly reduce customer loss, improve retention strategies, and optimize revenue. The system is scalable, automated, and extensible, making it suitable for real-world adoption in telecom, finance, SaaS, and subscription-based sectors. With future enhancements like real-time scoring, automated retention recommendations, and integration with CRM systems, this solution has the potential to evolve into a complete customer retention intelligence platform.

In conclusion, the project does more than build a predictive model — it enables business transformation. It converts raw customer data into strategic value, strengthens decision-making, and empowers organizations to retain customers before they churn, ultimately boosting profitability, customer loyalty, and sustainable growth.