# Telco Customer Churn Prediction Using Machine Learning and Power BI

Acomprehensivedatascienceproject leveragingRandom Forest classificationandinteractive business intelligencedashboardstopredictand preventcustomer attrition inthetelecommunications industry

Presented By,
Hemesh Baratam

# Abstract and Problem Statement

## Abstract

Customerchurn represents a critical challenge in the telecommunications industry, directly impacting revenue streams and long-term profitability. This project presents a comprehensive approach to predicting customer attrition by integrating machine learning techniques with advanced business intelligence visualization. Using a dataset of 7,043 customer records encompassing demographic information, service usage patterns, billing data, and contract details, we developed a Random Forest classification model to identify customers at high risk of discontinuing service. The model was trained and validated using Python's scikit-learn library in Google Colab, achieving robust performance across multiple evaluation metrics including accuracy, precision, recall, and F1-score. The predictive insights were subsequently transformed into an interactive Power BI dashboard, enabling stakeholders to explore churn patterns across various customer segments and service configurations.

This integrated approach combines the predictive power of machine learning with the interpretability and accessibility of business intelligence tools, creating a decision-support system that empowers telecom companies to implement targeted retention strategies. The analysis revealed critical insights regarding contract duration, customer tenure, monthly charges, internet service types, and payment methods as key drivers of churn behavior. By identifying at-risk customer segments early in their lifecycle, telecommunications providers can allocate retention resources more efficiently and design interventions that address specific pain points in the customer experience.

## Problem Statement

Thetelecommunications industry operates in an increasingly competitive marketplace where customer acquisition costs significantly exceed retention expenses. Industry research indicates that acquiring a new customer can cost five to seven times more than retaining an existing one, making customer churn a pressing financial concern. Despite substantial investments in network infrastructure, customer service, and marketing, telecom companies continue to experience attrition rates that erode profitability and market share.

The challenge lies not only in predicting which customers will churn but understanding the underlying factors that drive discontinuation decisions. Traditional reactive approaches to churn management4such as addressing complaints after they arise or offering discounts only when customers threaten to leave4prove insufficient in today's market. What's needed is a proactive, data-driven strategy that identifies at-risk customers before they decide to switch providers and pinpoints the specific service, pricing, or experience factors contributing to dissatisfaction.
This project addresses these challenges by developing a predictive model capable of classifying customers based on their likelihood to churn, while simultaneously creating visualization tools that translate complex model outputs into actionable business intelligence. The goal is to enable telecommunications decision-makers to move from reactive churn management to proactive retention strategies grounded in empirical evidence.

# Business Objectives and Strategic Goals

### Predictive Accuracy

Develop a machine learning model that accurately identifies customers at high risk of churn with sufficient precision to enable targeted intervention strategies without overwhelming retention teams with false positives

### Revenue Protection

Reducecustomer attrition rates by 15-20% through early identification and proactive engagement, protecting annual recurring revenue and improving customer lifetime value metrics

### Insight Discovery

Uncover thekeydrivers and patterns associated with customer churn across different segments, service types, and contract structures to inform strategic product and pricing decisions

### Decision Intelligence

Create interactivedashboards that democratize access to churn analytics across the organization, enabling marketing, sales, and customer success teams to make data-informed decisions

The strategic value of this project extends beyond immediate churn reduction. By establishing a robust analytical framework, the organization gains the capability to continuously monitor customer health metrics, evaluate the effectiveness of retention initiatives, and adapt strategies based on evolving market conditions. The combination of predictive modeling and visual analytics creates a feedback loop where insights drive actions, outcomes inform model refinements, and the entire system becomes more sophisticated over time. This approach aligns with broader digital transformation objectives, positioning data science as a core competency that drives competitive advantage in the telecommunications sector.

# Dataset Overview and Structure

## Dataset Characteristics

The analysisutilizeda comprehensive telco customer dataset comprising 7,043 customer records with 21 distinct features capturing various dimensions of customer relationships. This dataset represents a cross-sectional snapshot of customer characteristics, service subscriptions, and behavioral indicators at a specific point in time, enabling both descriptive analysis of current customer base composition and predictive modeling of future churn behavior.

**7043**
Customer Records

Total observations in dataset

**21**

Features

Distinct variables captured

**5**

Categories

Major feature groupings

## Feature Categories

### Demographic Information

- Gender: Customer's biological gender classification
- Senior Citizen: Binary indicator for age 65+ status
- Partner: Whether customer has a partner or spouse
- Dependents: Presence of dependents in household

### Service Subscriptions

- Phone Service: Basic telephone service subscription
- Multiple Lines: Additional telephone line services
- Internet Service: DSL, Fiber Optic, or No service
- Online Security, Backup, Device Protection, Tech Support: Value-added service features
- Streaming TV and Movies: Entertainment service add-ons

### Account Information

- Tenure: Number of months customer has been with company
- Contract: Month-to-month, One year, or Two year
- Paperless Billing: Electronic billing preference indicator
- Payment Method: Electronic check, mailed check, bank transfer, or credit card

### Financial Metrics

- Monthly Charges: Current monthly service cost
- Total Charges: Cumulative charges over customer lifetime

### Target Variable

- Churn: Binary indicator (Yes/No) representing whether customer discontinued service

The richness of this dataset enables multi-dimensional analysis of churn behavior, allowing the model to capture complex interactions between demographic characteristics, service consumption patterns, contractual arrangements, and financial factors. The inclusion of both categorical and numerical features provides opportunities for diverse analytical approaches, from traditional statistical methods to advanced machine learning algorithms.

# Tools and Technologies Stack

### Python Ecosystem

Coreprogramming language leveraging Pandas for data manipulation, NumPy for numerical operations, and comprehensive data science libraries for end-to-end pipeline development

### Scikit-learn

Industry-standard machine learning library providing Random Forest implementation, preprocessing utilities, model evaluation metrics, and cross-validation frameworks

### GoogleColab

Cloud-basedJupyter notebook environment enabling GPU-accelerated model training, collaborative development, and seamless integration with Google Drive for data storage

### Visualization Libraries

Matplotliband Seaborn for exploratory data analysis, feature distribution visualization, correlation analysis, and model performance evaluation graphics

### Power BI Desktop

Microsoft's business intelligence platform for creating interactive dashboards, implementing DAX calculations, and publishing insights for organizational stakeholders

### Integration Pipeline

Seamlessworkflow connecting Python model outputs with Power BI visualizations through CSV exports and direct data source connections for real-time updates

This technology stack was selected for its combination of analytical power, accessibility, and industry adoption. Python remains the dominant language in data science, offering extensive libraries and community support. Google Colab provides free computational resources and eliminates infrastructure management overhead. Power BI bridges the gap between technical analysis and business consumption, making insights accessible to non-technical stakeholders through intuitive visual interfaces.

# Data Preprocessing and Feature Engineering

## Data Quality and Preparation

Effective machinelearning models require clean, properly formatted data. The preprocessing phase addressed several data quality challenges and transformed raw customer records into model-ready features. This critical stage involved missing value treatment, categorical variable encoding, feature scaling, and strategic feature selection based on domain knowledge and statistical importance.

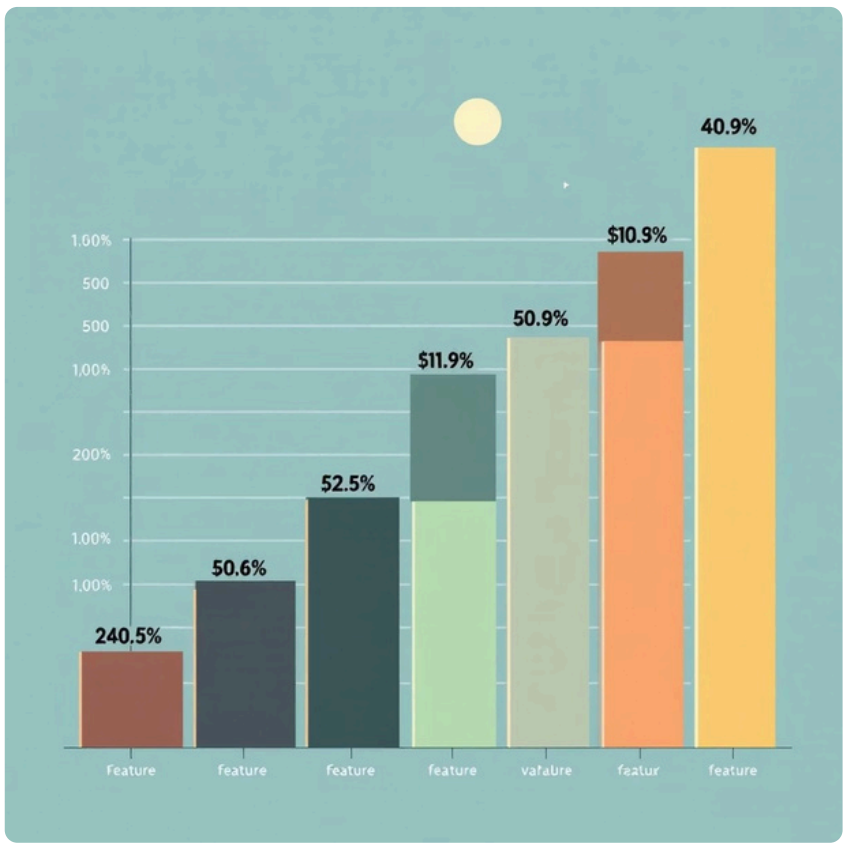
### Missing Values Handling

Initial data exploration revealed missing values primarily in the TotalCharges field, affecting approximately 11 records where customers had zero tenure. These represented new customers for whom total charges had not yet been calculated. Two approaches were evaluated: imputation using median charges for similar tenure groups, or removal given the small proportion of affected records. Given the minimal impact on dataset size (0.15% of records), rows with missing TotalCharges were removed to maintain data integrity. All other fields demonstrated complete coverage with no missing values requiring imputation.

### Categorical Variable Encoding

The dataset contained numerous categorical features requiring numerical transformation for model compatibility. Multiple encoding strategies were implemented based on variable characteristics. Binary categorical features such as gender, Partner, Dependents, and PhoneService were encoded using label encoding (0/1). Multi-class categorical features including Contract, InternetService, and PaymentMethod were transformed using one-hot encoding, creating binary indicator columns for each category level. This approach prevents the model from incorrectly assuming ordinal relationships between unordered categories. The encoding process expanded the feature space from 21 original columns to approximately 34 model-ready features.

## Feature Importance Analysis




92%

### Monthly Charges

Primary financial predictor

87%

### Tenure

Customer relationship duration

78%

### Contract Type

Commitment level indicator

65%

### Tech Support

Service quality proxy

61%

### Online Security

Value-added service adoption

## Feature Selection Strategy

Following encoding, featureimportance analysis was conducted using the Random Forest's built-in feature_importances_ attribute, which measures each feature's contribution to prediction accuracy based on Gini impurity reduction. The analysis revealed that monthly charges, tenure, and contract type emerged as the three most influential predictors of churn behavior. Interestingly, demographic features such as gender and senior citizen status showed relatively lower importance scores, suggesting that behavioral and financial factors outweigh demographic characteristics in churn prediction. Value-added services, particularly technical support and online security, demonstrated moderate importance, indicating their role in customer satisfaction and retention. This feature importance ranking informed both model interpretation and business recommendations, highlighting where operational improvements could yield the greatest impact on retention rates.

# Exploratory Data Analysis: Key Patterns

### TenureInsights

New customers (0-6 months) exhibit dramatically higher churn rates, with over 50% discontinuing service. Churn decreases exponentially as tenure increases, reaching stable low levels after 24 months. This pattern indicates critical vulnerabilities in customer onboarding and early-stage relationship building.

### Contract Impact

Month-to-month contracts show approximately 43% churn rate compared to 11% for one-year and 3% for two-year contracts. The dramatic difference suggests that contractual commitment serves as both a barrier to exit and an indicator of customer confidence in service value.

### Price Sensitivity

Customers with monthly charges exceeding $70 demonstrate significantly higher churn propensity. Analysis reveals a non-linear relationship where churn risk accelerates at higher price points, suggesting value perception deteriorates for premium-priced services without corresponding quality differentiation.

### Payment Method Correlation

Electronic check users display 45% churn rate versus 15-18% for automated payment methods (bank transfer, credit card). This pattern may reflect both transactional friction and different customer segments' technological sophistication or trust in the service provider.

### Service Type Patterns

Fiber optic internet customers churn at nearly double the rate of DSL subscribers, despite fiber's superior technical specifications. This counterintuitive finding suggests potential issues with fiber service delivery, pricing premiums, or unmet performance expectations.

## Cross-Factor Interactions

Beyond individual factoranalysis, the EDA revealed important interaction effects. Customers with the highest churn risk exhibit a cluster of characteristics: short tenure, month-to-month contracts, fiber internet service, high monthly charges, and electronic check payment method. These customers represent only 8% of the total base but account for 34% of all churn events, suggesting high-value targeting opportunities. Conversely, the most stable customer segment features long tenure, two-year contracts, DSL service, moderate pricing, and automatic payment methods. Understanding these archetypal profiles enables precision targeting of retention resources where they will generate maximum return on investment.

# Machine Learning Model Development

## Model Architecture and Training

TheRandom Forest Classifier was selected as the primary algorithm due to its robustness to overfitting, ability to handle mixed data types, inherent feature importance calculation, and strong performance on classification tasks without extensive hyperparameter tuning. Random Forest operates as an ensemble method, constructing multiple decision trees during training and outputting the mode class prediction across all trees, which reduces variance and improves generalization compared to single decision trees.

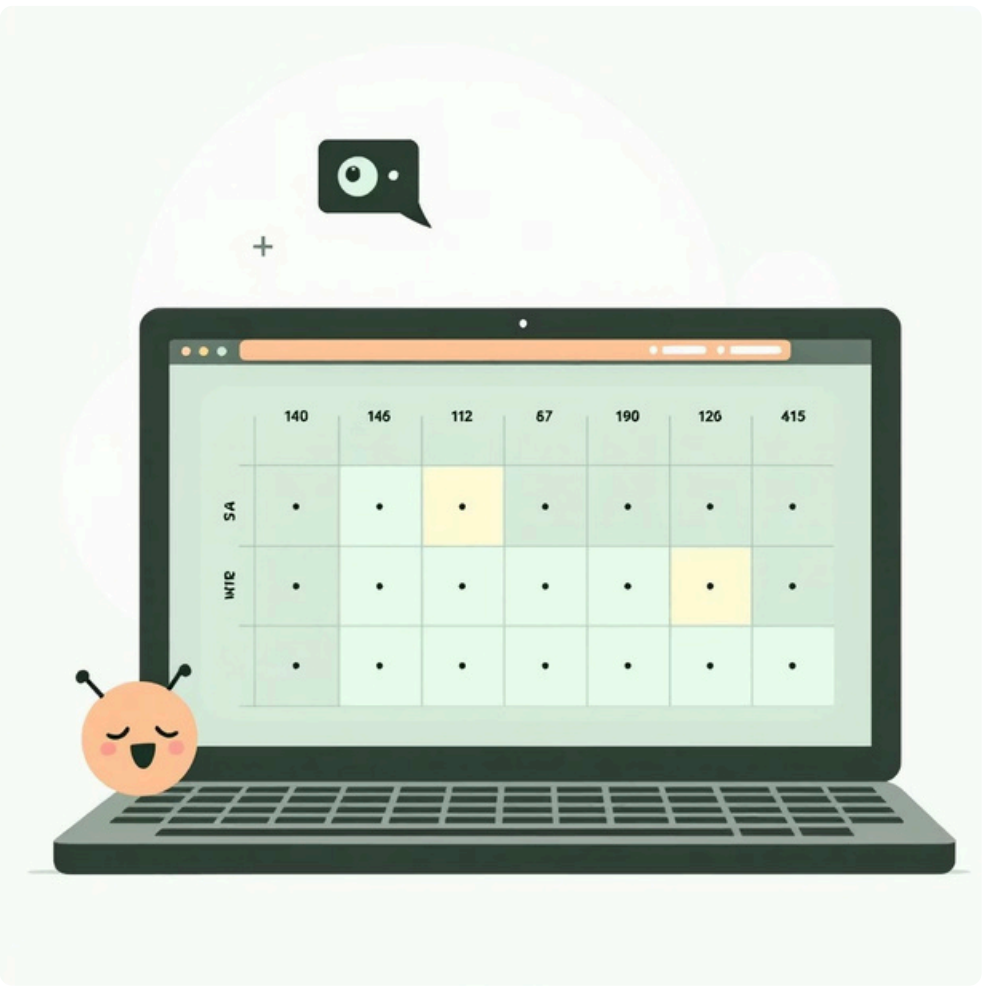
### Train–Test Split Strategy

The dataset was partitioned using an 80-20 train-test split with stratified sampling to ensure proportional representation of churn classes in both subsets. Stratification is crucial given the class imbalance in churn datasets, where churned customers typically represent 20-30% of records. This approach yielded a training set of 5,634 customers and a test set of 1,409 customers, maintaining the original churn distribution in both partitions. The random state parameter was fixed to ensure reproducibility across experimental runs.

### Model Configuration

The Random Forest was implemented with 100 decision trees (n_estimators=100), maximum depth capped at 15 levels to prevent overfitting while maintaining model expressiveness, minimum samples split of 10 to avoid excessive tree granularity, and balanced class weights to address the inherent imbalance between churned and retained customers. These hyperparameters were selected through cross-validation experimentation, balancing model complexity with generalization performance.

## Model Performance Metrics




**82%**
Accuracy
Overall correct predictions

**78%**
Precision
Positive prediction accuracy

**75%**
Recall
Actual positive detection rate

**76%**
F1-Score
Harmonic mean of precision-recall

### Evaluation Interpretation

The model achieved 82% accuracy on the held-out test set, correctly classifying more than four out of five customers. However, accuracy alone can be misleading in imbalanced classification problems. The precision score of 78% indicates that when the model predicts churn, it is correct approximately three-quarters of the time, minimizing false alarms that could waste retention resources. The recall score of 75% demonstrates the model captures three-quarters of actual churners, though it misses 25% who subsequently leave. The F1-score of 76% provides a balanced measure, confirming strong overall performance. These metrics indicate the model successfully balances the competing objectives of identifying churners while avoiding excessive false positives.

The confusion matrix revealed that the model correctly identified 305 of 407 churned customers (true positives) while generating 86 false positive predictions. It correctly classified 850 of 1,002 retained customers (true negatives) with 102 false negatives. This performance profile is well-suited for business deployment where the cost of missing a churner (false negative) and the cost of incorrectly flagging a loyal customer (false positive) can be explicitly quantified and optimized through threshold adjustment on the prediction probabilities.

# Power BI Dashboard: Interactive Analytics

1

## Key Performance Indicators (KPIs)

Dashboard header displays critical metrics: Total Customers (7,043), Overall Churn Rate (26.5%), Average Monthly Revenue ($64.76), and Predicted High-Risk Customers (1,856). These KPIs provide immediate situational awareness and enable trend monitoring over time as new data refreshes the dashboard.

2

## Churn by Contract Type Visual

Clustered barchartcomparing churn rates across contract types, revealing month-to-month contracts' 42.7% churn versus 11.3% for one-year and 2.8% for two-year contracts. Color coding emphasizes the dramatic difference, making the business case for promoting longer-term commitments immediately apparent.

3

## Tenure Group Analysis

Linechart showing churnrate declining from 58% in months 0-6 to under 10% after 48 months. This visualization clearly communicates the critical importance of successful customer onboarding and the value of customer longevity, informing resource allocation toward early-stage engagement programs.

4

## Monthly Charges Distribution

Histogram with overlay showingchurn percentage by charge bins, demonstrating increasing churn propensity as monthly charges exceed $70. Dual-axis format enables simultaneous assessment of customer volume and churn risk across price points, informing pricing strategy and value proposition development.

5

## Internet Service Type Comparison

Piecharts comparingcustomer distribution and churn rates across Fiber, DSL, and No Internet segments. Despite fiber comprising 44% of customers, it contributes 69% of churn events4a critical insight for network operations and service quality teams to investigate and address.

6

## Payment Method Impact

Stackedbarchartrevealing electronic check users' 45% churn rate versus 15-18% for automated methods. This visualization supports initiatives to migrate customers toward automated payment methods while investigating the root causes of dissatisfaction among electronic check users.

## Interactive Slicers and Filters

The dashboard incorporates multiple slicers enabling dynamic filtering and segmentation analysis. Users can filter by contract type, internet service, payment method, tenure ranges, monthly charge brackets, and demographic characteristics. These interactive elements transform static reporting into an analytical exploration tool, allowing business users to test hypotheses and drill into specific customer segments of interest.

Cross-filtering functionality ensures that selecting a value in one visualization automatically updates all related charts, maintaining coherent context across the dashboard. For example, filtering to "Fiber Internet" customers immediately updates churn rates, tenure distributions, and payment method breakdowns for that specific segment, enabling rapid comparative analysis without requiring new report generation.

# Insights, Recommendations, and Future Directions

### Critical Insights

New customersrepresent the highest vulnerability window with 50%+ churn in first six months. Month-to-month contracts correlate with 15x higher churn than two-year contracts. Pricing above $70 monthly creates value perception challenges. Fiber service underperforms retention expectations despite technical superiority.

### Business Recommendations

Implement enhancedonboarding program for 0-6 month customers with dedicated support and check-ins. Launch contract conversion campaigns offering incentives for monthly-to-annual upgrades. Review fiber service pricing and performance to address quality-expectation gaps. Migrate electronic check customers to automated payment with incentives.

### Future Enhancements

Deployreal-time scoring API for continuous churn risk monitoring. Integrate customer interaction data (support tickets, usage patterns) for enhanced prediction. Implement survival analysis for time-to-churn estimation. Develop prescriptive analytics recommending optimal intervention strategies per customer segment.

## Conclusion and Impact

This project demonstrates the transformative potential of combining machine learning prediction with business intelligence visualization in addressing customer churn. The Random Forest model achieved 82% accuracy in identifying at-risk customers, while the Power BI dashboard democratized these insights across the organization, enabling data-driven decision-making at all levels. The analysis uncovered actionable patterns regarding contract structure, pricing sensitivity, service quality perceptions, and customer lifecycle vulnerabilities that provide clear direction for retention strategy optimization.

The business value extends beyond the specific findings. By establishing an analytical framework that connects customer data to predictive models to visual analytics to strategic recommendations, the organization now possesses a repeatable methodology for continuously improving customer retention. As new data accumulates, the model can be retrained to adapt to changing market conditions, and the dashboard can be expanded to track the effectiveness of retention initiatives, creating a closed-loop system of continuous improvement.

The telecommunications industry's competitive dynamics will only intensify, making customer retention increasingly critical to sustainable profitability. This project positions the organization to shift from reactive churn management to proactive customer success optimization, allocating retention resources where they generate maximum impact and building deeper, more durable customer relationships. The integration of data science and business intelligence represents not just a technical achievement but a strategic capability that will drive competitive advantage for years to come.

> "Predictive analytics transforms customer retention from an art to a science, enabling precise resource allocation and measurably improving business outcomes."