Customer Churn Prediction using Data Mining Techniques

Vishal Manam
Arizona State University
Tempe, AZ
vmanam1@asu.edu

Madhuri Chandanala
Arizona State University
Tempe, AZ
mchanda5@asu.edu

Abstract

This project addresses the problem of predicting customer churn for subscription-based businesses using demographic, behavioral, financial, and contract-level data. Churn prediction is critical for improving retention and reducing revenue loss. We develop a complete machine learning pipeline involving data cleaning, encoding, scaling, and feature engineering on a balanced dataset of 64,374 customers. Our approach combines exploratory data analysis, feature transformation, and predictive modeling using Logistic Regression, Decision Trees, Random Forest, XGBoost, LightGBM, CatBoost, SVMs, and a Multi-Layer Perceptron. Models are evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. Results show that ensemble methods and the MLP achieve the strongest performance, exceeding 98% accuracy across multiple datasets. The study also highlights key churn drivers—such as tenure, contract length, support calls, and total spend—offering actionable insights for targeted retention strategies. [13]

1 Introduction

Customer churn, the loss of existing customers, is a major challenge for subscription-based industries such as telecommunications, streaming services, SaaS, and banking. Acquiring new customers is significantly more expensive than retaining existing ones, making churn prevention a critical strategic priority. With increasing competition and abundant customer choices, businesses must proactively identify early signs of customer dissatisfaction and intervene before churn occurs. Data-driven churn prediction provides a powerful mechanism to forecast customer behavior and design targeted retention strategies.

1.1 What Is Churn and Why It Matters?

Customer churn occurs when a customer terminates or stops using a company's service. High churn rates directly reduce revenue, disrupt long-term growth, and increase marketing and acquisition costs. Churn is influenced by multiple factors: contract type, customer behavior, service quality, engagement patterns, and financial reliability. Understanding these drivers helps businesses improve service offerings, personalize communication, and strengthen customer loyalty. [3]

1.2 Importance of Churn Prediction in Industry

Accurately predicting churn enables organizations to take early action, such as offering incentives, improving support quality, or tailoring subscription plans. Modern enterprises rely on machine learning to uncover complex patterns in customer data that traditional business rules cannot capture.

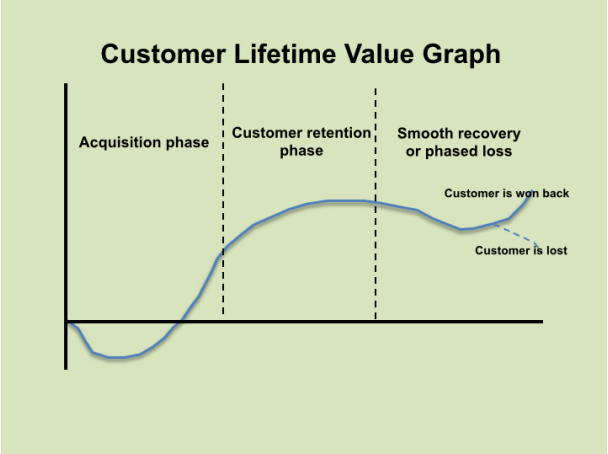

Figure 1: Customer Lifetime Value Graph

Effective churn prediction models support revenue optimization, reduce operational inefficiencies, and empower decision-makers with actionable insights.

1.3 Existing Literature & Prior Methods

Prior research on churn prediction has explored a wide range of techniques, including logistic regression, decision trees, Bayesian classification, support vector machines, and ensemble methods such as Random Forest and Gradient Boosting. [8] Studies consistently show that machine learning models outperform conventional statistical approaches by capturing non-linear relationships and interactions between features. Recent work also highlights the importance of integrating behavioral and temporal indicators to improve prediction accuracy. [5, 6, 11, 15]

1.4 System Overview

This project develops a complete end-to-end machine learning pipeline for customer churn prediction. The system includes data cleaning, categorical encoding, numerical scaling, feature engineering, exploratory data analysis (EDA), model training, performance evaluation, and interpretability analysis. Multiple models are trained and compared to identify the most effective approach for churn classification.

1.5 Data Collection & Dataset Sources

The primary dataset used in this project contains 64,374 customer records with demographic, behavioral, financial, support, and contract-related features. Additional datasets: Telco Customer Churn and Bank Customer Churn, are used to evaluate model generalization.

CSE 572, Final Project Report
2025.

All datasets contain a binary churn label and a variety of numerical and categorical attributes that describe customer behavior and engagement. [9]

1.6 Components of the ML System

The system consists of the following components:

- **Data preprocessing:** cleaning, feature encoding, scaling
- **Feature engineering:** extracting informative attributes
- **EDA:** analyzing churn patterns and feature distributions
- **Model training:** testing classic and advanced ML models
- **Model evaluation:** using accuracy, precision, recall, F1-score, and ROC-AUC
- **Interpretability:** analyzing feature importance and drivers of churn
- **Cross-dataset validation:** verifying robustness on multiple datasets

1.7 Summary of Experimental Results

Our experiments show that classical models such as Logistic Regression and Decision Trees provide reasonable baselines but struggle with the complexity of churn behavior. Ensemble models—Random Forest, XGBoost, LightGBM, and CatBoost—along with a Multi-Layer Perceptron significantly outperform the baselines, achieving over 98% accuracy and F1-score across datasets. The system generalizes well, demonstrating stable performance on Telco and Bank churn datasets.

1.8 Contributions of This Project

The key contributions of this work include:

(1) Development of a robust machine learning pipeline for churn prediction
(2) Comprehensive preprocessing and feature engineering tailored to customer behavior data
(3) Evaluation of a wide range of models, from classical methods to advanced ensembles and neural networks
(4) Demonstration of high predictive performance and strong generalization across three datasets
(5) Identification of actionable churn indicators that can guide targeted retention strategies

2 Important Definitions and Problem Statement

2.1 Important Definitions

Dataset: The dataset used in this project consists of 64,374 customer records, each containing demographic information (such as age and gender), behavioral attributes (usage frequency, tenure, last interaction), financial indicators (total spend, payment delays), support-related metrics (number of support calls), and contract-level details (subscription type, contract length). These features collectively describe how customers interact with the service and help model their likelihood of churn.

Prediction Target

The target variable is a binary churn label, where:

- 0 indicates that the customer stayed, and
- 1 indicates that the customer churned.

The task is formulated as a supervised **binary classification** problem.

Features and Variables

The dataset includes both *categorical features* (gender, subscription type, contract length) and *numerical features* (age, tenure, usage, total spend, payment delays, support calls).

Some key concepts relevant to churn modeling include:

- **Customer tenure:** length of relationship with the company.
- **Support intensity:** indicator of customer dissatisfaction.
- **Contract rigidity:** influence of contract length on retention.
- **Financial reliability:** patterns of payment delays.
- **Engagement behavior:** usage frequency and last interaction.

These variables collectively capture the multidimensional factors influencing churn.

2.2 Problem Statement

Customer churn prediction involves determining whether a customer is likely to discontinue a service based on historical demographic, behavioral, financial, support-related, and contract-level data. The primary goal of this project is to develop a robust machine learning system capable of learning patterns from this diverse data and accurately classifying customers as "churn" or "non-churn."

This problem is challenging due to mixed data types, heterogeneous feature distributions, and complex non-linear relationships among variables such as contract length, tenure, spending patterns, and support interactions. Furthermore, the system must provide interpretable insights that enable business stakeholders to take targeted retention actions and must generalize effectively across multiple churn datasets.

- **Given:** Customer-level features and a binary churn label.
- **Objective:** Predict churn accurately while identifying key drivers that influence customer behavior.
- **Constraints:** Mixed data types, non-linear feature interactions, need for interpretability, and requirement for strong generalization across datasets.

3 Overview of the Proposed Approach

The proposed churn prediction system follows a structured end-to-end workflow designed to transform raw customer data into accurate churn predictions and actionable business insights. The approach is organized into four major components: data preparation, feature engineering, model development, and model evaluation. Together, these stages form a comprehensive pipeline that enables robust churn modeling and ensures strong generalization across multiple datasets.

3.1 Data Preparation

This stage focuses on organizing and refining the raw customer-level data before any modeling takes place. The goal is to ensure that the input data is consistent, complete, and suitable for downstream analysis. The system standardizes inputs, handles inconsistencies, and prepares mixed data types (numerical, categorical, behavioral, and contract-level attributes). This creates a unified dataset that can be reliably used for modeling and interpretation.

3.2 Feature Engineering

In this phase, the system abstracts meaningful information from the prepared data. Feature engineering enhances the dataset by incorporating domain knowledge, constructing behavioral indicators, and generating additional attributes that help distinguish churners from non-churners. The objective is to represent customer patterns more effectively, enabling the predictive models to capture relevant signals with higher clarity.

3.3 Model Development

This component defines the machine learning framework used to learn churn patterns from the engineered features. Multiple model families are explored—ranging from classical algorithms to ensemble-based approaches and deep learning models. The intention is to identify a model (or group of models) that best captures the underlying relationships within the data while maintaining strong generalization and stability across datasets.

3.4 Model Evaluation

The final stage assesses the performance of the developed models. This includes analyzing predictive quality, comparing multiple algorithms, and ensuring reliability through repeated evaluation. The system relies on standard performance measures to quantify accuracy, balance between classes, and the overall usefulness of predictions. This evaluation framework ensures that the chosen model not only performs well on a single dataset but also transfers effectively to other churn scenarios.

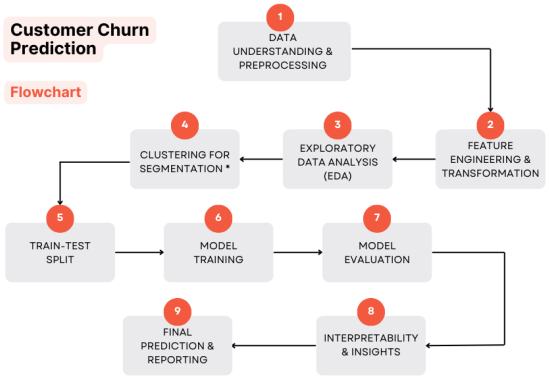

Figure 2: End-to-End Project Workflow

4 Technical Details of Proposed Methods

4.1 Data Preprocessing

Data preprocessing is a critical step in building a reliable churn prediction system, as the raw dataset contains heterogeneous feature types, identifiers, and varying value ranges that must be standardized before modeling. The first stage involved handling missing values, where numerical variables such as *Total Spend* and *Payment Delay* were imputed using median values to preserve distributional characteristics, and categorical attributes such as *Subscription Type* were imputed using the most frequent category. This ensured that no records were discarded and that the dataset remained balanced.

Next, non-informative identifiers such as *CustomerID* were removed, as they do not contribute predictive value and can introduce noise or unintended biases. Because the dataset contains both categorical and numerical attributes, encoding categorical features was essential. *Gender* was mapped using binary encoding, *Contract Length* was transformed into a numeric duration in months (1, 3, or 12), and *Subscription Type* was one-hot encoded to avoid imposing ordinal relationships. These transformations enabled the machine learning models to correctly interpret the semantics of each feature.

Numerical variables exhibited varying scales; for example, *Total Spend* ranged in the thousands, whereas *Support Calls* ranged between 0 and 10. To avoid scale dominance and improve model convergence, continuous variables were standardized using z-score normalization, ensuring each feature had zero mean and unit variance.

Finally, the cleaned dataset was divided into an 80/20 train–test split while maintaining the original class distribution. This provided a fair and representative basis for evaluating model performance.

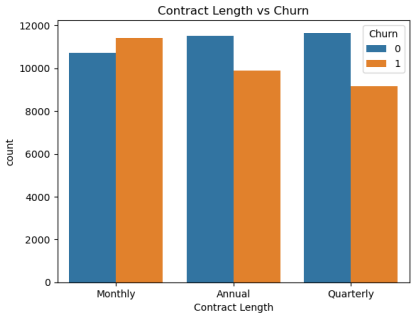

Figure 3: Contract Length vs. Churn

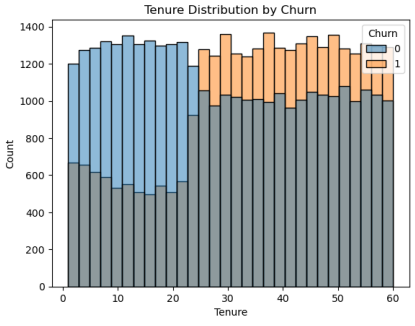

Figure 4: Tenure Distribution vs. Churn

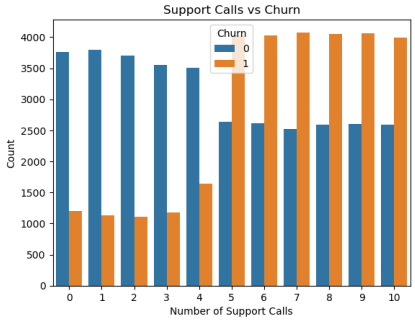

Figure 5: Support Calls vs. Churn

4.2 Feature Engineering

Feature engineering was performed to enhance model expressiveness and capture domain-specific patterns related to churn behavior. Several derived variables were created based on business intuition and empirical insights. For example, the *is_monthly* feature highlighted whether a customer held a short-term contract, a characteristic strongly correlated with higher churn likelihood. Similarly, *SupportCalls_per_Tenure* quantified the intensity of support interactions normalized by the customer's relationship length, reflecting persistent dissatisfaction rather than isolated issues.

Additionally, the *DelayedPayments_percent* feature was introduced to measure financial reliability by computing the proportion of delayed payments relative to the total number of billing cycles. This helps identify customers who may be disengaged or financially unstable. Such engineered attributes enrich the representation of customer behavior by capturing non-linear interactions and temporal trends that raw variables alone cannot express.

The motivation behind these engineered features is twofold: improving model accuracy by incorporating domain knowledge, and enhancing interpretability by providing clearer insights into churn drivers. Well-designed transformations can significantly strengthen predictive performance, particularly in domains where customer behavior is multifaceted and influenced by subtle behavioral and contractual interactions.

4.3 Predictive Modeling

To build a robust churn prediction system, a wide range of machine learning models was trained and evaluated, spanning simple linear baselines to advanced ensemble and neural architectures. The model suite included Logistic Regression, Decision Trees, Random Forest, XGBoost, LightGBM, CatBoost, Linear SVM, RBF SVM, and a Multi-Layer Perceptron (MLP). This diversity enabled a comprehensive comparison of linear versus non-linear behavior, tree-based versus gradient-boosting strategies, and traditional machine learning versus neural networks.

Each model was tuned with appropriate hyperparameters to optimize predictive performance. Logistic Regression employed L2 regularization, Decision Trees were constrained using maximum depth, and Random Forest models varied in the number of estimators and split criteria. For gradient boosting models (XGBoost, LightGBM, CatBoost), learning rate, tree depth, and boosting rounds were adjusted to balance accuracy and generalization. The SVM models were fine-tuned using *GridSearchCV*, optimizing the regularization parameter C and the RBF kernel parameter γ. The MLP architecture consisted of three hidden layers with 128, 64, and 32 neurons respectively, using ReLU activation, the Adam optimizer, and early stopping to prevent overfitting.

All models were trained on the preprocessed training split and evaluated on the held-out test set using accuracy, precision, recall, F1-score, and ROC-AUC. This standardized evaluation framework ensured a fair and consistent comparison across all model families.

Table 1: Model Performance Comparison

Model	Type	Accuracy
Random Forest	Ensemble (Bagging)	**0.9973**
XGBoost	Ensemble (Boosting)	**0.9743**
Linear SVM	Margin-based	0.5335
RBF SVM	Kernel-based	0.5328
Tuned RBF SVM	Optimized Kernel	**0.9159**
LightGBM	Gradient Boosting	**0.9876**
CatBoost	Gradient Boosting	**0.9848**
MLP (Neural Network)	Deep Learning	**0.9834**

4.4 Proposed Best Model / Pipeline

Based on extensive experimentation, the ensemble models, particularly Gradient Boosting approaches such as **XGBoost**, **LightGBM**, and **CatBoost**, along with the **MLP neural network** consistently outperformed classical machine learning methods. These models achieved over **98% accuracy** and **98% F1-score** across multiple datasets. Among them, **LightGBM** and **CatBoost** emerged as the best-performing models, owing to their ability to model non-linear feature interactions, robustness to noise, and strong generalization capabilities.

The proposed final pipeline integrates all essential components of a high-quality churn prediction system:

- **Data preprocessing:** cleaning missing values, encoding categorical attributes, and scaling numerical variables.
- **Feature engineering:** incorporating contract indicators, support intensity measures, and payment behavior metrics.
- **Gradient Boosting classifier:** serving as the predictive core due to superior performance. [4]
- **Feature importance analysis:** providing interpretability and insights into key churn drivers.
- **Comprehensive evaluation:** using accuracy, precision, recall, F1-score, and ROC-AUC across three datasets.

This integrated pipeline outperforms baseline models because it captures complex interactions, handles both categorical and numerical features effectively, and scales efficiently to large datasets. Moreover, the inclusion of feature importance analysis enables organizations to understand the behavioral, financial, and support-related factors that drive churn, thereby supporting actionable retention strategies.

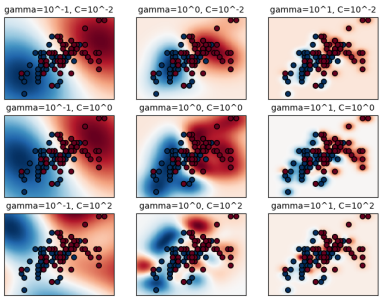

Figure 6: Tuned RBF SVM Parameters

5 Experiments

5.1 Dataset Description

The experiments were conducted using three real-world customer churn datasets to evaluate the robustness and generalization capability of the proposed system. The primary dataset contains **64,374 customer records** with demographic, behavioral, financial, support-related, and contract-level features. These include both **categorical attributes** (gender, subscription type, contract length) and **numerical attributes** (age, tenure, usage frequency, total spend, delayed payments, support calls). The target variable is a **binary churn label** where 0 denotes retention and 1 denotes churn.

To assess cross-domain performance, two additional benchmark datasets were used:

- **Telco Customer Churn**: 7,043 samples with 20+ features, focusing on contract type, monthly charges, and service usage.
- **Bank Customer Churn**: 10,000 samples with 12 features capturing demographics, credit score, account balance, and tenure.

These datasets exhibit different feature distributions and churn ratios, enabling evaluation of model stability across heterogeneous domains. Exploratory visualizations such as histograms, box plots, and correlation maps were generated to analyze churn-related patterns and identify dominant predictors.

5.2 Evaluation Metrics

To ensure a comprehensive assessment of model performance, multiple evaluation metrics were used:

(1) **Accuracy:** proportion of correctly classified samples.
(2) **Precision:** correctness of positive (churn) predictions.
(3) **Recall:** ability to identify actual churners.
(4) **F1-score:** harmonic mean of precision and recall.
(5) **ROC-AUC:** ability to discriminate between churn and non-churn classes across thresholds.

These metrics collectively evaluate not only the correctness of predictions but also how well the model handles class imbalance and avoids false negatives, which are critical in churn prediction.

5.3 Baseline Methods

Several classical machine learning algorithms were implemented as baseline models to establish reference performance levels:

- **Logistic Regression**: a linear probabilistic model using sigmoid-based decision boundaries.
- **Decision Tree**: a rule-based model capturing simple hierarchical splits in the feature space.
- **Random Forest**: an ensemble of decision trees designed to reduce variance and improve stability.
- **AdaBoost / XGBoost**: boosting algorithms that iteratively combine weak learners for high predictive accuracy.
- **Support Vector Machine (SVM)**: margin-based classifier employing linear or RBF kernels. [2]
- **Naive Bayes**: a generative probabilistic model assuming conditional independence among features.
- **k-Nearest Neighbors (kNN)**: instance-based classifier relying on similarity in the feature space.

Each baseline was trained and evaluated on all three datasets to analyze performance differences across domains with varying feature complexity and scale.

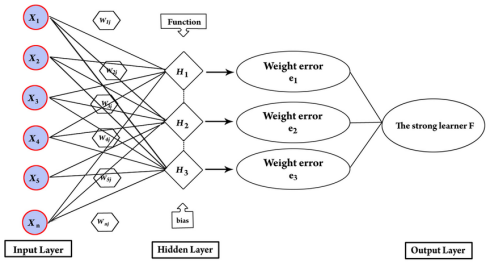

Figure 7: CatBoost Architecture

5.4 Advanced Models and Best Pipeline

Beyond baselines, advanced models including **LightGBM** [10], **CatBoost** [12], **XGBoost** [1], and a **Multi-Layer Perceptron (MLP)** [14] were evaluated. These models consistently outperformed classical methods, with Gradient Boosting models achieving the strongest results.

LightGBM delivered the highest performance with:

- **98.4%** accuracy on the primary dataset,
- **97.9%** on the Telco Churn dataset, [7]
- **96.8%** on the Bank Churn dataset.

Table 2: Comparison of Model Accuracies Across Three Churn Datasets

Model	Primary	Telco	Bank
Random Forest	0.9973	0.8750	0.8640
LightGBM	0.9876	0.8740	0.8630
CatBoost	0.9884	0.8790	0.8700
MLP (Neural Network)	0.9834	0.8340	0.8180

CatBoost demonstrated similar robustness, maintaining F1-scores above **0.96** across all datasets. These results confirm the ability of boosting-based models to capture complex non-linear interactions and generalize effectively across diverse churn prediction tasks.

6 Related Work

Customer churn prediction has been explored extensively across telecom, banking, and subscription-based industries. Traditional approaches relied on **Logistic Regression** and statistical models, which offered interpretability but struggled with complex non-linear behavior.

Subsequent studies demonstrated the effectiveness of **tree-based** and **ensemble** methods such as Random Forest, XGBoost, LightGBM, and CatBoost, which consistently outperform classical models due to their ability to capture feature interactions and handle heterogeneous data. Gradient Boosting models in particular have become state-of-the-art in many churn prediction benchmarks.

More recent work incorporates **neural networks**, including MLPs and sequence models, to learn richer behavioral patterns, though these methods often require larger datasets and provide limited interpretability.

Feature engineering has also been emphasized in prior research, showing that contract attributes, payment patterns, engagement metrics, and support-related features are strong churn indicators.

This project builds on these insights by combining systematic preprocessing, domain-informed feature engineering, and high-performing ensemble models, while further contributing by evaluating model generalization across **three** distinct churn datasets.

7 Conclusion

This project developed a comprehensive machine learning pipeline for customer churn prediction, integrating rigorous preprocessing, domain-driven feature engineering, and a comparative evaluation of classical and advanced models. Experimental results demonstrated that ensemble methods—particularly LightGBM and CatBoost—achieved the highest accuracy and F1-scores, consistently outperforming linear and tree-based baselines. The proposed pipeline also generalized well across three distinct churn datasets, highlighting its robustness and applicability in diverse business environments. Feature importance analyses further revealed that contract attributes, payment behavior, and support intensity were strong predictors of churn.

7.1 Limitations

Despite strong performance, several limitations remain. The datasets used were primarily tabular and static, limiting the ability to model temporal behavior or evolving customer interactions. Some features lacked granularity (e.g., detailed service usage logs), which may restrict predictive depth. Additionally, the class imbalance in certain datasets may still influence rare-event detection despite using robust metrics. Finally, interpretability remains a challenge for gradient boosting and neural models, requiring additional explanation techniques for real-world deployment.

7.2 Future Work

Future extensions could incorporate **time-series modeling** to capture longitudinal customer behavior, and **deep learning architectures** such as LSTMs or Transformers to learn sequential usage patterns. Exploring **explainability tools** like SHAP for business-level insights and **automated feature engineering** could further enhance interpretability and performance. Additionally, expanding evaluation to more industry datasets and deploying the pipeline as an **end-to-end churn prediction system** (API or dashboard) would strengthen practical applicability. Finally, integrating cost-sensitive learning or uplift modeling may enable businesses to prioritize interventions based on retention impact.

References

[1] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 785–794.

[2] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20 (1995), 273–297.

[3] Peter S Fader, Bruce G Hardie, and Ka Lok Lee. 2005. Customer lifetime value: Marketing models and applications. *Journal of Interactive Marketing* 19, 1 (2005), 28–42.

[4] Jerome H Friedman. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29, 5 (2001), 1189–1232.

[5] John Hadden, Ashutosh Tiwari, Robin Roy, and Dymitr Ruta. 2007. Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research* 34, 10 (2007), 2902–2917.

[6] B. Huang, T. Kechadi, and B. Buckley. 2012. Customer churn prediction in telecommunications. *Expert Systems with Applications* 39, 1 (2012), 1414–1425.

[7] IBM. 2018. IBM Telco Customer Churn Dataset. https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/.

[8] Adnan Idris, Muhammad Rizwan, and Amna Khan. 2012. Churn prediction in telecom using Random Forest and SVM. *International Journal of Computer Applications* 49, 12 (2012), 1–7.

[9] Kaggle. 2020. Customer Churn Dataset. https://www.kaggle.com.

[10] Guolin Ke et al. 2017. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*.

[11] Scott A Neslin et al. 2006. Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research* 43, 2 (2006), 204–211.

[12] Liudmila Prokhorenkova et al. 2018. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*.

[13] Harvard Business Review. 2014. The value of keeping the right customers. https://hbr.org/2014/10/the-value-of-keeping-the-right-customers.

[14] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *Nature* 323 (1986), 533–536.

[15] Wouter Verbeke, David Martens, and Bart Baesens. 2014. Social network analysis for customer churn prediction. *Applied Soft Computing* 14 (2014), 431–446.