Using Data Analysis to Predict and Reduce Customer Churn in the Telecommunications Industry

Executive Summary/ Abstract

Customer churn is a major challenge in the telecommunication industry because losing customers can reduce company revenue and increase the cost of acquiring new customers.This report investigates how data analytics can be used to predict customer churn and support business decision-making. A telecom customer churn dataset obtained from Kaggle was analysed using python in Google Colab. The study involves data cleaning, handling missing values, and preparing the dataset for analysis. Linear regression was used to examine factors influencing customer spending, while logistic regression was used to predict whether customers were likely to leave the company. In addition, A/B testing was evaluated as a method for validating customer retention strategies. The findings suggest that data-driven approaches can help telecom companies identify high-risk customers, improve retention efforts, and make more informed business decisions.

Introduction

In today's business environment, data has become an important resource for supporting decision-making and improving organisational performance. Companies collect large amounts of customer data every day and use this information to understand customer behaviour, improve services, and increase profitability. Data analytics helps organisations transform raw data into useful insights that can support strategic and operational decisions. As a result,

businesses are increasingly investing in analytical tools and machine learning techniques to gain a competitive advantage.

One of the major challenges faced by the telecommunications industry is customer churn.
Customer churn occurs when customers stop using a company's services and switch to a competitor. High churn rates can reduce revenue, increase marketing expenses, and negatively affect long-term business growth. research has shown that retaining existing customers is often more cost-effective than acquiring new ones, making customer retention an important business objective (Ahmad, Jafar and Aljoumma, 2109)

This report investigates how data analytics can be used to predict customer churn in the
telecommunication sector. The analysis is based on a Telecom Customer churn dataset obtained from Kaggle and processed using Python in Google Colab. The study explores preprocessing techniques, regression analysis, classification methods, and validation approaches such as A/B testing. In addition, relevant academic literature is reviewed to critically examine how predictive analytics can support customer retention strategies.
The aim of this report is to demonstrate how data-driven insights can help telecommunication companies identify customers at risk of leaving and make more informed business decisions.

Business Problem Identification

Customer churn is one of the most significant challenges faced by telecommunications companies. Customer churn occurs when customers stop using a company's services and move to a competing provider. In a highly competitive telecom market, losing customers can lead to reduced revenue, lower profitability, and increased marketing costs because attracting new customers is often more expensive than retaining existing ones.

The dataset used in this project contains customer information such as contract type, tenure, internet services, monthly charges, and churn status. These

variables provide an opportunity to analyse customer behavior and identify factors that may influence a customer's decision to leave the company.

According to previous research, customer churn remains a major concern in the telecommunications industry, and predictive analytics can help organisations identify customers who are at risk of leaving before the actual churn occurs (Ahmad, Jafar, and Aljoumma, 2019). Therefore, the business problem addressed in this report is how a telecom company can use customer data and machine learning techniques to predict churn
and support customer retention strategies. Solving this problem can help businesses improve customer loyalty, reduce financial losses, and make more effective data-driven decisions.

Data acquisition and preprocessing

The dataset used in this study was obtained from a publicly available customer churn dataset and analysed using Google Colab. The purpose of data acquisition was to collect relevant customer information that could help predict whether a customer would leave a service provider. The dataset contained demographic information, account details, service subscriptions, and customer behavior attributes that are commonly associated with churn prediction.

Before model development, it was necessary to perform data preprocessing because raw datasets often contain inconsistencies, missing values, and data formats that are not suitable for machine learning algorithms. An initial exploratory analysis was conducted to understand the structure of the dataset, identify data types, and examine the distribution of variables. This step helped reveal potential data quality issues that could negatively affect prediction performance.

The preprocessing stage involved several important tasks. First, irrelevant identifiers
such as customer IDs were removed because they do not contribute meaningful information for churn prediction. Second, missing values were examined and handled appropriately to ensure that incomplete records would not introduce bias

into the analysis. Third, categorical variables such as gender, contact type, and internet service options were converted into numerical representations through encoding techniques. This transformation was necessary because machine learning algorithms require numerical input data.

Feature scaling was also considered for variables with different value ranges. Standardisation and normalisation techniques are widely recommended in machine learning literature because they help prevent variables with larger scales from dominating the learning process (Geron,2022). In addition, duplicate records and inconsistent entries were checked and removed where necessary to improve overall data quality.

A further preprocessing step involved splitting the dataset into training and testing subsets. The training set was used to develop the machine learning models, while the testing set was reserved for evaluating predictive performance on unseen data. This approach reduces the risk of overfitting and provides a more reliable assessment of model effectiveness (Han, Kamber and Pei, 2002).

Overall, data acquisition and preprocessing formed a critical foundation for the study. Careful preparation of the dataset ensured that the subsequent machine learning models were trained on accurate, consistent , and meaningful data, thereby improving the reliability of churn prediction results.

Regression Analysis

As part of my customer churn prediction project, I decided to include regression analysis to understand which factors most influence customer churn. Unlike classification, which simply predicts whether a customer will leave or stay, regression helps me see the strength and direction of relationships between different variables (Huang et al., 2012).

I used logistic regression because my target variable is binary - churn is either "yes" or "no". This is a common starting point for beginners like me, as it is easy to interpret. From the telecom dataset I downloaded from Kaggle, I looked at

features such as monthly charges, total charges, contract type, tenure, and payment method.

My regression results showed that tenure (how long a customer stays) has a strong negative relationship with churn. In simple words: the longer a customer stays, the less likely they are to leave. This makes practical sense because long-term customers probably have built habits or loyalty. Monthly charges also showed a positive relationship, customers paying higher monthly fees tended to churn more.

However, I must be critical here. Logistic regression has limitations. As Ahmad et al.(2019) explain in their study on SyriaTel data, tree-based algorithms like XGBoost performed much better than regression for churn prediction, achieving 93.3% AUC compared to lower results from simpler models. Similarly, the literature review by Imani (2025) highlights that regression struggles with imbalanced datasets , and my dataset had only about 5-15% churn cases, which is very unbalanced.

Therefore, while regression helped me understand why churn happens (e.g., high monthly charges, short tenure), I realised it is not the best tool for accurate prediction. I will use it more for explanation than for final prediction.

Classification analysis

After completing regression analysis, I moved to classification, which is more commonly used in churn prediction studies. Unlike regression that gives me relationships, classification puts customers into groups, in my case, "will churn" or "will not churn".
This is exactly what telecom companies need: a clear warning about which customers are at risk.

Following the approach used by Ahmad et al. (2019) on SyriaTel data, I tested several classification algorithms on my Kaggle dataset. I started with Decision Tree because it is easy to visualise and explain to non-technical people. Then I used Random forest, which combines many decision trees to reduce errors. Finally, I applied XGBoost, which the literature consistently praises as the best performer for churn prediction.

My results showed that XGBoost performed best, similar to what Ahmad et al. (2019) found with their 93.3% AUC score. Random Forest came second, and simple Decision Tree was third.However, I must be honest, my accuracy was not as high as the published research. I think this is because I did not use Social Network analysis (SNA) features, which the SyriaTel study added to improve results from 84% to 93.3%.

A critical lesson I learned is about class imbalance. My dataset had very few churn cases, maybe only 5% to 10% of all customers. As Imani (2025) explains in his literature review, most machine learning algorithms become biased toward the majority class (non-churn) and fail to catch churners properly. I tried undersampling, but I lost useful data. In hindsight, I should have used SMOTE or other techniques mentioned by Wagh and Wagh (2022)

Despite these struggles, Classification gave me a practical model. I now understand that XGBoost is worth learning properly, and handling imbalance is not optional, it is essential.

Hypothesis Testing and A/B Testing

I initially thought hypothesis testing would be very useful for the churn project. The idea is simple: I make an assumption (null hypothesis) and then test whether my data supports it or not. For example, I tested whether customers with month-to-month contracts churn more than those with yearly contracts. M null hypothesis was "there is no difference". The result? I rejected it, month to month customers really do churn more.

However, I must be honest about a limitation I discovered. As I read through the journal articles provided, including Ahmad et al. (2019) and literature review by Imani (2025), I noticed that none of these published churn prediction studies used traditional hypothesis testing. They all focused on machine learning AUC scores, not p-values or t-tests.

Similarly, A/B testing, where a company tests two strategies on different customer groups, was not applicable to my project. Why? Because A/B testing requires live experiments with real customers, and I only had a historical dataset from Kaggle. I cannot send half my customers a retention offer and see what happens. That would require working with an actual telecom company.

So what is my conclusion here? Hypothesis testing and A/B testing are valuable in business contexts, but for a secondary data analysis project like mine, they are not practical. My "testing" happened through train-test splits and cross-validation, not through statistical hypothesis tests.

Findings

From my analysis of the telecom customer churn dataset using Google Colab, I identified five key findings. These are based on my regression and classification work, supported by the journal article I reviewed.

Finding 1: Long-term customers are less likely to leave

My regression analysis clearly showed a negative relationship between tenure and churn.
Customers who stayed with the company for two or more years were much less likely to churn compared to new customers. This makes business sense, longer tenure means the customer has already invested time and probably built habits. Ahmad et al. (2019) also found tenure to be a strong predictor in their SyriaTel study.

Finding 2: Higher monthly charges increase churn risk

I found that customers paying higher monthly charges showed greater churn probability.

This was surprising to me at first because I thought higher-paying customers might be more satisfied. But the data suggests the opposite, they feel they are paying too much and may find cheaper alternatives elsewhere. Wagh and Wagh (2022) similarly identified pricing as a key churn driver in their research.

Finding 3: Month-to-month contracts show higher churn

This was one of my clearest findings. Customers on month-to-month contracts churned significantly more than those on one-year or two-year contracts. The reason seems obvious, once I thought about it, longer contracts create a lock-in effect. Customers cannot easily leave without penalty, so they stay.
Month-to-month customers have no such barrier.

Finding 4: Retention offers may improve loyalty

Although my dataset did not directly include retention offers, I can infer from the findings above that targeted offers could help. For example, offering a discount to switch from month-to-month to a yearly contract, or reducing monthly charges for high-risk customers. Imani (2025) notes in his literature review that retention strategies based on churn prediction are widely used in the telecom industry.

Finding 5: Classification models can identify high-risk customers

My classification analysis showed that XGBoost performed best at predicting which customers would churn. While my accuracy was not as high as Ahmad et al. (2019) who achieved 93.3% AUC, I was still able to identify a clear high-risk group. The practical implication is that telecom companies do not need to offer retention deals to everyone, they can focus resources only on those most likely to leave.

Critical reflection

I must be honest, my findings are not perfect. My dataset was imbalanced, with very few churn cases. This means my model may have missed some churners. Also, I did not have social network features like the SyriaTel study used. Despite these limitations, my five findings are consistent with the published literature and provide useful direction for a telecom company wanting to reduce churn.

Recommendations

Based on my five findings, I now propose five practical recommendations for a telecom company wanting to reduce customer churn. These are written from a business perspective, not just technical.

Recommendation 1: Develop customer retention programs

My finding that month-to-month contracts show higher churn tells me that retention programs should target these customers first. For example, offering a small discount to switch to a one-year contract could reduce churn significantly. Ahmad et al. (2019) showed that predicting churn is useless without action, prediction must lead to retention strategies. I agree completely.

Recommendation 2: Target high-risk customers using predictive models

My classification analysis proved that SGBoost can identify which customers are likely to churn. Instead of wasting money on offers for everyone, the company should use my model to score all customers weekly and send retention offers only to the top 10-20% highest-risk customers. This is efficient and data-driven. However, I must add that the model needs retraining every few months because customers' behaviour changes over time, as Ahmad et al. (2019) also noted.

Recommendation 3: Offer loyalty rewards for long-term customers

My first findings showed that long-term customers are less likely to churn, but that does not mean the company should ignore them. Loyalty rewards, such as free upgrades, discounts, or priority support, can reinforce their staying behaviour. Wagh and Wagh (2022) argue that keeping existing customers is cheaper than finding new ones. I think this is a simple but powerful business truth.

Conclusion

This project aimed to predict customer churn for a telecom company using machine learning. I downloaded a dataset from Kaggle, performed data cleaning in Google Colab,
and applied regression and classification techniques. My goal was not just to build a model but to understand what drives churn and how a business can respond.

From my analysis, I found five key things. Long-term customers stay. High monthly charges push customers away. Month-to-month contracts are risky. Retention offers can help. And classification models like XGBoost can successfully identify high-risk customers. These findings are consistent with the published literature. I reviewed, including Ahmad et al. (2019), Wagh and Wagh (2022), and Imani (2025).

However, I must be honest about my limitations. My dataset was imbalanced, with very few churn cases. I did not have access to social network features, which the SyriaTel study used to reach 93.3% AUC. My own accuracy was lower. Also, I could not perform A/B testing because I only had historical data.

Despite these challenges, I believe my project achieves its aim. I built a working churn prediction model, identified practical findings, and proposed realistic recommendations.

For a first-time, as a student, I am satisfied with what I learned. If I had more time, I would focus on handling class imbalance better and testing my model on newer data. But for now, this project shows that machine learning can help telecom companies reduce churn and keep more customers.

Dataset Citation:

https://www.kaggle.com/code/usmandatalab/telecom-customer-churn-analysis

Google Colab:

https://colab.research.google.com/drive/1gevB4AIWdhXN4ZOyN6FZgCO_jtgOG_MJ#scrollTo=XDoT59ktqGrc

References (Harvard style)

Ahmad, A.K., Jafar, A. and Aljoumma, K. (2019) 'Customer churn prediction in telecom using machine learning in big data platform, Journal of Big Fata, 6(1), p. 28.

Huang, B.Q., Kechadi, M.T. and Buckly, B. (2012) 'Customer churn prediction in telecommunication', Expert systems with Applications, 39(1), pp. 1414-1425.

Imani, M. (2025) 'Customer churn prediction in telecommunication industry: a literature review', preprints.org, doi:10.20944/preprints202403.0585.v3.

Geron, A. (2022) Hands-on Machine Learning with Scikit-Learn, Kera and TensorFlow. 3rd edn. Sebastopol: 0'Reilly Media.
Han, J., Kamber, M. and Pei, J. (2022) Data Mining: Concepts and Techniques. 4th edn.
Burlington: Morgan Kaufmann.

Wagh, S.K. and Wagh, K.S. (2022) 'Customer churn prediction in telecom sector using machine learning techniques', SSRN, doi:10.2139/ssrn.4158415.

Industry Report:

From:

- IBM
- McKinsey & Company
- Deloitte

Academic Articles: Source-Google Scholar

1. **Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models**
2. **Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models**
3. **Implementing machine learning techniques for customer retention and churn prediction in telecommunications**
4. **Customer Churn Prediction in Telecommunication Industry Using Deep Learning**
5. **Predictive Analytics for Telecom Customer Churn: Enhancing Retention Strategies in the US Market**

By sitting this assessment, I am confirming:

That I have worked independently on this assessment submission, and I have not worked together with any current or previous student at the University to produce my submission, other than when officially permitted to do so;
· I also confirm the contents of my submission have not been generated by a third party.;

· I have fully referenced and correctly cited the work of others, where required;

· I have not used any generative AI tools to generate, rephrase or otherwise produce content for this submission, except where their use was explicitly permitted, and I have read and understood the University's AI in Higher Education Policy and Protocols;
· I have read the Student Discipline Regulations and understand that any academic misconduct can lead to disciplinary consequences and undermine academic integrity;

· I understand that where applicable, I am expected to engage with my academic work in a manner that meets the professional standards and requirements set by the relevant Professional, Statutory and Regulatory Body (PSRB) or accrediting body for my course.
By submitting this assessment submission, I am confirming that I am fit to sit according to the Assessment Regulations.