FINAL PROJECT

TELCO CHURN ANALYSIS REPORT

Contents

Introduction:

Customer loyalty is a critical factor in ensuring stable and sustained revenue for any business. It refers to a long-term commitment in which customers repeatedly purchase a company's products or services. Organizations value loyal customers not only for their consistent financial contribution but also because they often serve as informal brand ambassadors. However, customer loyalty is not constant; it may fluctuate due to competitive pressures, variations in product or service quality, shifting brand preferences, or other situational influences.

Understanding the causes behind declining loyalty is essential for effective business management. When customers cease purchasing a company's products or discontinue using their services, the phenomenon is known as **customer churn**. A rising churn rate signals underlying issues that may be deteriorating the customer experience or diminishing perceived value. Customer loss results in reduced revenue and may also harm the organization's reputation and long-term growth prospects.

To address this challenge, businesses must identify the root causes of churn by analyzing customer trends, feedback, service quality concerns, and behavioral or engagement patterns. The **churn rate**, a key performance metric, provides insight into the proportion of customers who discontinue their relationship with a business over a specified period.

Customer churn—also referred to as customer defection—is particularly prevalent in industries such as telecommunications, where competition is high and service differentiation is limited. In this sector, churn significantly influences a company's competitiveness and market position. Telecom providers routinely use churn prediction models to estimate how many customers are likely to leave and to design proactive retention strategies aimed at sustaining revenue and attracting new subscribers.

The objective of this project is to develop and implement a Business Intelligence (BI) system capable of predicting customer churn, identifying the factors contributing to customer departure, and recommending data-driven interventions to minimize churn.

For this project I have selected the "Telco Customer Churn" dataset from kaggle.com. This dataset contains 7043 records of 23 features based on geographical location, family pattern, telecom service subscriptions, and method of payments.

Executive Summary

Customer churn presents a significant risk to revenue stability and long-term competitiveness in the telecommunications industry. This **Business Intelligence (BI)** project analyzes customer churn using a California-based telecom dataset containing 7,043 customer records and 23 attributes related to demographics, service usage, billing behavior, and customer value. The objective of the study is to identify the key drivers of churn, develop predictive models to detect at-risk customers, and provide actionable, data-driven recommendations to improve customer retention.

Exploratory analysis reveals a churn rate of 26.5%, indicating that more than one in four customers discontinue service within the observed period as statistical testing shows that churn is strongly associated with contract type, service offerings, billing methods, and pricing. Customers on month-to-month contracts, those without value-added services such as Tech Support or Online Security, and customers with higher monthly charges exhibit substantially higher churn risk. Tenure and Customer Lifetime Value (CLTV) are negatively correlated with churn, highlighting the importance of long-term engagement and customer value retention.

Two supervised machine learning models were developed and evaluated: Logistic Regression and Random Forest. Logistic Regression achieved an accuracy of 74% and demonstrated strong recall for churners, making it effective for identifying at-risk customers. The Random Forest model outperformed Logistic Regression with an accuracy of 81%, producing fewer false positives and offering better precision in identifying customers likely to stay. Moreover, feature importance analysis confirms that tenure, contract type, monthly charges, and payment methods are the most influential predictors of churn.

Geographical analysis further indicates that churn is concentrated in large and competitive urban markets such as Los Angeles, San Diego, and San Francisco, suggesting the need for localized retention strategies. Based on these insights, the project recommends service bundling, pricing optimization, targeted retention campaigns for high-risk segments, and a strategic emphasis on transitioning customers to long-term contracts. Overall, this BI solution demonstrates how predictive analytics can support proactive churn management and informed decision-making in the telecom sector.

Problem

Customer churn is a critical challenge for telecommunications companies, as losing subscribers directly impacts revenue, profitability, and long-term sustainability. Understanding why customers leave is essential to developing effective retention strategies. Despite having extensive customer data, telecom companies often struggle to identify the key drivers of churn and to implement data-driven solutions that target high-risk segments. In this report, I focus on three specific areas of concern that are central to the churn problem.

First, operational and service-related factors can strongly influence a customer's decision to leave. Services like Tech Support, Online Security, Online Backup, and the type of contract may affect customer satisfaction. Customers who face frequent service issues or feel their plan is not suitable may be more likely to churn. Finding out which of these factors have the biggest impact can help the company improve services and keep more customers.

Second, pricing and customer value are important in understanding churn. High monthly charges or poor value for money can push customers to switch to other providers. Customers with high Customer Lifetime Value (CLTV) are especially important, as losing them can lead to large financial losses. Understanding the link between monthly charges, CLTV, and churn can help the company adjust prices, offer better deals, and focus on retaining its most valuable customers.

Finally, geographic factors also influence churn, as customer behavior may vary across cities, states, or regions. Certain locations may show higher churn rates due to differences in competition, service availability, or socio-economic factors. Identifying cities or regions with the highest churn allows for targeted interventions, such as localized marketing, tailored service plans, or infrastructure investments to improve customer retention.

Overall, these problems show the need for a thorough analysis of churn. By looking at service issues, pricing, customer value, and location, the company can reduce churn, retain valuable customers, and stay competitive.

Solution

Feature Dropping and Data Cleaning

During data cleaning, several columns were removed because they functioned only as identifiers, contained excessive missing values, or offered little predictive value for churn modelling. **CustomerID** was excluded as it was purely an identifier, while **Count** was constant across all records. **Country** and **State** were dropped due to having only a single value, providing no variability. **Churn Label** was removed because it duplicated the target variable (**Churn Value**). **Churn Reason** contained values only for churned customers, resulting in substantial missing data, and **Churn Score**, a precomputed metric from IBM SPSS, was excluded to prevent data leakage.

Additionally, several location-based variables (**City, Zip Code, Latitude, Longitude, and Lat Long**) had very high cardinality, which could increase dimensionality and lead to overfitting. To address this, these columns were removed and replaced with an engineered feature, **Zip3**, created by extracting the first three digits of the ZIP code. This approach reduced the number of unique categories while preserving meaningful regional patterns relevant to customer churn.

Final Dataset after Cleaning and Feature Engineering

After data cleaning and feature engineering, the dataset contains **7,043 records and 22 features**, including **18 categorical variables**, **three numerical variables** (Tenure Months, Monthly Charges, CLTV), and **one target variable** (Churn Value). All records are complete with no missing or duplicate values.

The features represent customer demographics, service subscriptions, contract and billing preferences, financial behavior, and a regional indicator (**Zip3**). This streamlined and well-structured dataset supports effective churn prediction while maintaining interpretability and minimizing complexity.

Chi-Square test of independence

Ran a **Chi-Square test of independence** between each **categorical feature** and the **target variable (Churn Value)**, to identify which categorical variables have a **statistically significant relationship** with churn.

Out of the 17 categorical features analyzed — *Gender, Senior Citizen, Partner, Dependents, Phone Service, Multiple Lines, Internet Service, Online Security, Online Backup, Device Protection, Tech Support, Streaming TV, Streaming Movies, Contract, Paperless Billing, Payment Method,* and *Zip3* — **14 exhibited statistically significant associations** ($p < 0.05$) with customer churn. In contrast, features such as **Gender, Phone Service,** and **Zip3** showed no significant relationship with churn, indicating that customer attrition is largely consistent across these demographic and regional categories.

Point-Biserial Correlation Analysis

Point-biserial correlation was used to assess the relationship between numerical features and the binary target variable (Churn Value). All correlations were statistically significant ($p < 0.05$). **Tenure Months** showed the strongest negative correlation ($r = -0.35$), indicating that longer-tenured customers are less likely to churn. **Monthly Charges** displayed a positive correlation ($r = 0.19$), suggesting higher costs increase churn risk, while **Total Charges** and **CLTV** were negatively correlated, reflecting greater loyalty among higher-value customers.

Based on these results, the model uses **14 categorical features** (including contract, service, and billing attributes) and **four numerical features** (Tenure Months, Monthly Charges, Total Charges, and CLTV). This combination captures both behavioral and financial drivers of churn, resulting in a compact and effective dataset suitable for supervised models such as Logistic Regression and Random Forests.

Results From Logistic Regression:

The model achieves an **overall accuracy of 74%** and performs well in identifying non-churn customers (Precision = 0.90, Recall = 0.73). For churners, it captures **78% of actual churn cases**, though with lower precision (0.51), indicating some false positives. This trade-off is acceptable in churn modelling, where identifying potential churners is prioritized over precision.

Coefficient analysis shows that **Tenure Months, fixed-term contracts, and Online Security** significantly reduce churn risk, while **higher Monthly Charges, Fiber Optic internet, and**

Electronic Check payments increase the likelihood of churn. Overall, the results indicate that pricing, contract structure, and payment methods drive churn, whereas long-term engagement and value-added services improve customer retention.

Results From Random Forest Classifier:

The model achieved an **accuracy of 81%**, outperforming Logistic Regression (74%). It performs **very well at identifying non-churners**, with a **Recall of 0.90** and **Precision of 0.85**, meaning it's highly reliable in recognizing customers who will stay. For **churners**, it correctly identifies **55%** (Recall = 0.55), with **Precision = 0.66**, suggesting some false negatives — customers who actually churn but were predicted as non-churn.

Overall, the model balances bias and variance better than Logistic Regression, showing stronger generalization and more robust predictive accuracy.

Model Comparison:

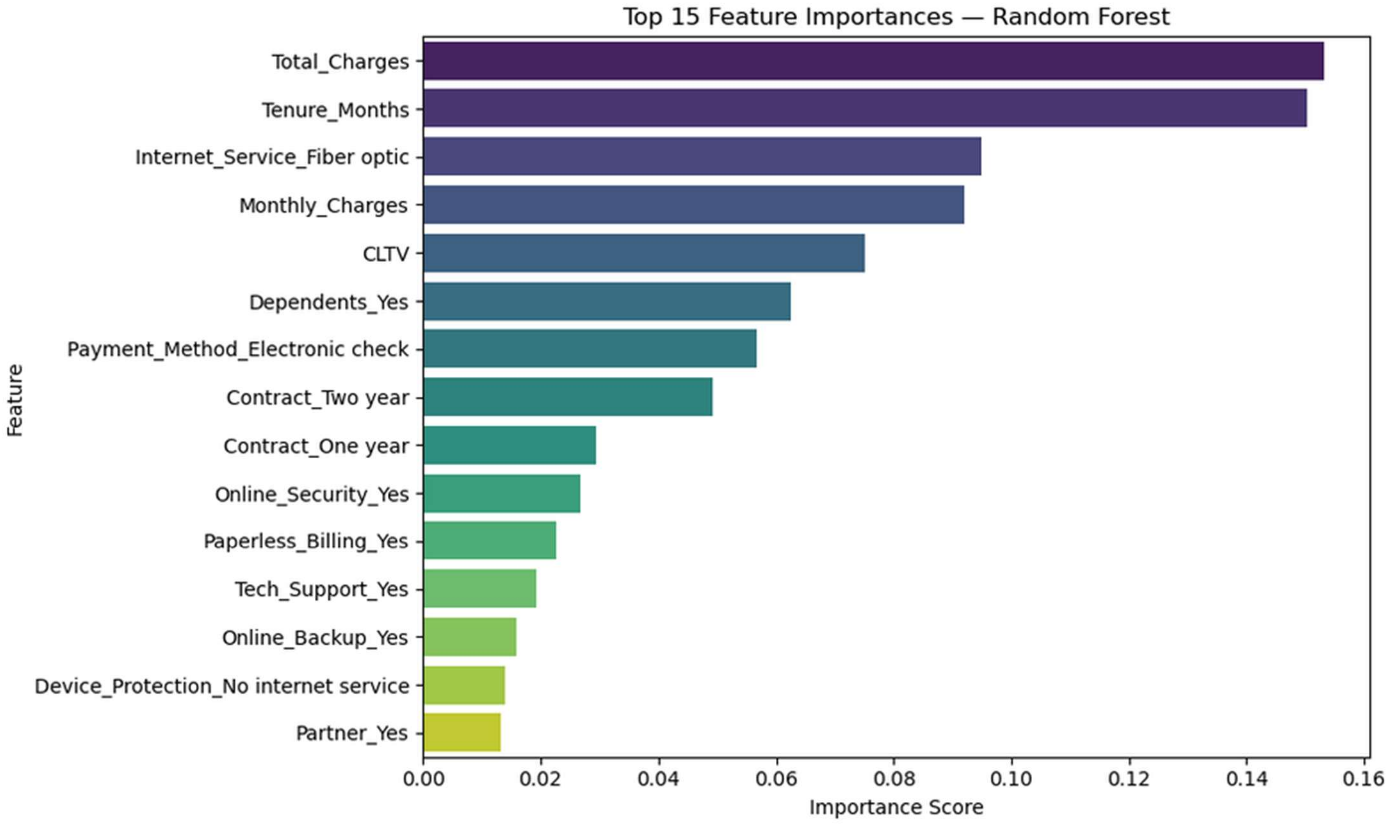
Model	**Actual: No Churn Predicted: No Churn (TN)**	**Actual: No Churn Predicted: Churn (FP)**	**Actual: Churn Predicted: No Churn (FN)**	**Actual: Churn Predicted: Churn (TP)**
Logistic Regression	757	278	84	290
Random Forest	**930**	**105**	**169**	**205**

The Random Forest model correctly identifies more non-churn customers (**TN = 930**) than Logistic Regression (**TN = 757**), making it more conservative and accurate in predicting customer retention. It also produces fewer false positives (**105 vs. 278**), reducing unnecessary retention efforts. However, this conservatism results in more missed churners (**FN = 169**) compared to Logistic Regression (**FN = 84**).

Logistic Regression, on the other hand, correctly identifies more churners (**TP = 290 vs. 205**), achieving higher recall and making it more effective at detecting at-risk customers, though at the cost of lower precision.

From a business perspective, **Logistic Regression** is better suited for proactive churn prevention when minimizing missed churners is critical, while **Random Forest** is preferable for cost-efficient, targeted retention strategies that prioritize precision and resource optimization.

Feature Importance Analysis (Random Forest Classifier)



The Random Forest model provides feature importance scores, which indicate each variable's contribution to predicting churn. The Random Forest feature importance analysis reveals that customer longevity (Tenure, Total Charges) and billing characteristics (Monthly Charges, Contract Type, and Payment Method) are the strongest churn determinants.

Key Findings:

Customer churn poses a significant challenge to long-term profitability in the telecommunications industry. This study analyses a California-based telecom dataset comprising **7,043 customers** to identify the primary drivers of churn and assess their operational, financial, and geographic impact. The overall churn rate in the dataset is **26.5% (1,869 customers)**, indicating that more than one in four customers discontinue service, underscoring the importance of proactive retention strategies.

Operational and service-related factors emerge as the strongest drivers of churn. Customers on **month-to-month contracts account for 55.0% (3,875 customers)** of the customer base and exhibit the highest churn risk due to low switching barriers. Additionally, nearly half of all customers lack key value-added services: **49.3% do not have Tech Support** and **49.7% do not have Online Security**. The absence of these services increases customer dissatisfaction and reduces service stickiness. Furthermore, **Fiber optic customers (44.0%)**, who typically pay higher monthly fees, demonstrate increased churn sensitivity when service expectations are unmet.

Pricing and customer value analysis reveals a strong relationship between higher costs and churn. Customers who churn pay an average monthly charge of **$74.44**, compared to **$61.27** for retained customers, representing a **21.5% difference**. The median monthly charge among churned customers (**$79.65**) further confirms that churn is concentrated in higher-priced plans. From a revenue perspective, the **average Customer Lifetime Value (CLTV) is $4,400**, with the **top 20% of customers (1,411 customers)** contributing **26.91% of total revenue**. Although high-CLTV customers churn less frequently (**20.62%**) than lower-value customers (**28.02%**), their churn results in disproportionately high long-term revenue loss.

Geographic analysis shows that churn is not evenly distributed across locations. While customers are spread across **1,129 cities and 1,652 ZIP codes**, churn is heavily concentrated in large, competitive urban markets. **Los Angeles (90 churned customers)** and **San Diego (50 churned customers)** account for the highest churn volumes, followed by **San Francisco (31)**, **San Jose (29)**, and **Sacramento (26)**. Several mid-sized cities, including **Fresno (16)** and **Long Beach (15)**, also exhibit notable churn levels, suggesting localized service or pricing challenges.

Challenges and Ethical Considerations

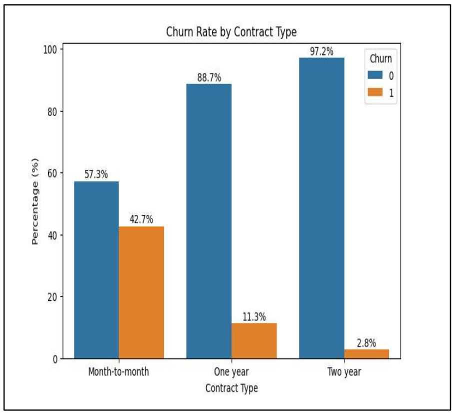
The **Telco Customer Churn analysis** has several inherent **data and modelling limitations** that may influence the interpretation and generalizability of the results. The dataset covers only a single fiscal quarter and represents customers from one region, California, which limits the ability to generalize findings across different geographies or time periods. Another limitation arises from **class imbalance**, as only about 26% of the records represent churned customers. This imbalance can bias the model toward predicting "No Churn," reducing sensitivity to actual churners. Certain features, such as **Churn Score**, were pre-engineered or derived variables. They may introduce **data leakage**, since they can indirectly encode churn-related information that the model should not have access to. Lastly, the models used, **Logistic Regression** and **Random Forest,** come with their own constraints. Logistic Regression assumes a linear relationship between features and churn probability, which may oversimplify complex interactions, while Random Forest, if not properly tuned, can risk **overfitting**, especially on a limited dataset.

From an **ethical standpoint**, several key considerations must guide the responsible use of churn prediction models. **Customer privacy** can be compromised, as telecom data often contains sensitive personal and billing information. Organizations must ensure strict **data anonymization** and adhere to data protection regulations. Another crucial factor is **fairness and bias**. Features like *Senior Citizen* or *Zip3* (regional clusters) could serve as **indirect proxies for demographic or socioeconomic bias**, potentially leading to discriminatory outcomes if predictions are used unfairly. Therefore, regular bias audits and fairness checks should accompany any model deployment. Moreover, **transparency** is essential for models like Random Forest, which deliver strong accuracy; their ensemble structure can hide how decisions are made. Techniques such as **feature importance visualization** should be used to explain predictions in an interpretable manner. The **responsible use** of predictive insights is equally critical: churn models should be leveraged to improve customer satisfaction and enhance service quality and not exclude or penalize specific customer groups. Finally, strong **data governance** practices must ensure that customer data is collected with informed consent, securely stored, and retained only as long as necessary to justify its analytical purpose.

Business Recommendations

Based on our analysis, several actions can help reduce customer churn Following recommendations can be proposed:

1. **Service Improvements:**
   Customers *without Tech Support or Online Security* are more likely to churn. The company should promote these services as essential rather than optional by *offering* ***affordable packages and highlighting their value*** in marketing. This builds customer *confidence, increases perceived value, and encourages longer-term contracts*.
2. **Pricing Adjustments:**
   Pricing strongly affects churn, especially for customers with *high monthly charges and low CLTV*. The company should target this high-risk group with ***personalized discounts or bundled services***, such as combining internet, streaming, and device protection at a lower price. This increases perceived value, improves retention, and maintains profitability.



3. **Cities with highest churn rate**
   For cities with high churn, the company should apply ***targeted retention strategies*** such as ***early engagement***, ***improved customer support***, and special service bundles. Offering onboarding assistance and incentives to move from *month to month to* ***longer contracts*** can improve satisfaction. Focusing on high-churn cities ensures resources are used efficiently and retention efforts are tailored where they are needed most.

Finally, promoting *long-term contracts combined with loyalty rewards and pricing improvements* can strengthen customer retention and stabilize revenue.

Applicability of the BI Solution to Other Sectors

The proposed Business Intelligence framework is highly transferable to industries with recurring revenue models. Its modular pipeline—comprising data preprocessing, feature engineering, statistical validation, supervised learning, and model interpretation—can be adapted across domains with minimal modification.

In **banking and financial services**, the framework can predict customer attrition using transaction behavior, account tenure, and product utilization. In **insurance**, it supports non-renewal risk detection based on premium levels, claims history, and policy duration. For **e-commerce and retail**, the approach models repeat purchase likelihood using pricing sensitivity, loyalty engagement, and fulfilment metrics. In **healthcare and wellness platforms**, it enables early detection of user disengagement through service utilization and billing patterns. Similarly, **streaming and SaaS** providers can forecast subscription cancellations using usage intensity, plan attributes, pricing, and support interactions.

Overall, this BI solution provides a scalable, industry-agnostic churn prediction architecture that enables proactive, data-driven retention strategies.

Conclusion

This project demonstrates the effectiveness of Business Intelligence and predictive analytics in identifying the key drivers of customer churn. By integrating statistical analysis with machine learning models, the study provides actionable insights into how pricing, contract structure, service usage, and customer tenure influence attrition. The results support proactive, data-driven retention strategies and highlight the broader applicability of the BI framework across subscription-based industries.

References

[1] S. Macko, “Telco Customer Churn in IBM Cognos Analytics 11.1.3,” *IBM Community Blogs*, Jul. 11, 2019.

[2] T. Amin, A. Khan, and M. Qamar, “Predicting Customer Churn in Telecom Industry Using Machine Learning Techniques,” *2020 IEEE 23rd International Multitopic Conference (INMIC)*, Bahawalpur, Pakistan, 2020, pp. 1–6. doi: 10.1109/INMIC50486.2020.9318112

[3] G. Verbeke, D. Martens, C. Mues, and B. Baesens, “Building comprehensible customer churn prediction models with advanced rule induction techniques,” *Expert Systems with Applications*, vol. 38, no. 3, pp. 2354–2364, Mar. 2011. doi: 10.1016/j.eswa.2010.08.023

[4] S. Ahmed, F. Khan, and A. Alenezi, “Customer churn prediction in telecom using machine learning in big data platform,” *IEEE Access*, vol. 8, pp. 213606–213619, 2020. doi: 10.1109/ACCESS.2020.3039341

[5] J. A. Larson and R. De Carvalho, “Ethical Machine Learning for Predictive Analytics: Challenges and Solutions,” *2021 IEEE International Conference on Big Data (Big Data)*, Orlando, FL, USA, 2021, pp. 5422–5431. doi: 10.1109/BigData52589.2021.9671479

[6] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” *arXiv preprint arXiv:1702.08608*, 2017.