## **PROJECT REPORT**

**Project Title: Customer Churn Analysis for Telecommunication Company**

**Team Members:**

Akshay Belnekar

Rohit Sangem

Vaibhavi Hedaoo

## **ORIGINAL WORK STATEMENT**

We the undersigned certify that the actual composition of this proposal was done by us and is original work.

| **Sr No.** | **Typed Name** | **Signature** |
| --- | --- | --- |
| 1. | Akshay Belnekar | AB |
| 2. | Rohit Sangem | RS |
| 3. | Vaibhavi Hedaoo | VH |

**I. Executive Summary**

In our study, we delved into understanding why customers leave a telecommunications company, employing PySpark for an in-depth data analysis. Our investigation aimed to unearth the key reasons behind customer churn, offering valuable insights for the company to enhance customer retention strategies.

We uncovered that higher monthly charges and shorter tenure were significant predictors of churn. Customers who pay more each month or are newer to the service tend to leave at higher rates. This finding underscores the importance of pricing and initial customer experience in retaining customers. Moreover, our analysis revealed that the source through which customers learn about the company also influences their decision to stay or leave. For instance, customers acquired through merchant collaborations showed a higher tendency to churn, suggesting a possible mismatch in expectations or satisfaction.

Our research stands out by providing a clear narrative on the dynamics of customer churn, supported by data. We offer actionable insights for the company, suggesting a reevaluation of pricing strategies and partner collaborations, alongside enhancing the onboarding experience for new customers. The goal is to not just curb churn but to foster a more satisfying and enduring relationship with customers, ensuring they find value and satisfaction in the company's offerings. Our recommendations aim to transform these insights into strategic actions, enabling the company to improve customer loyalty and reduce churn effectively.

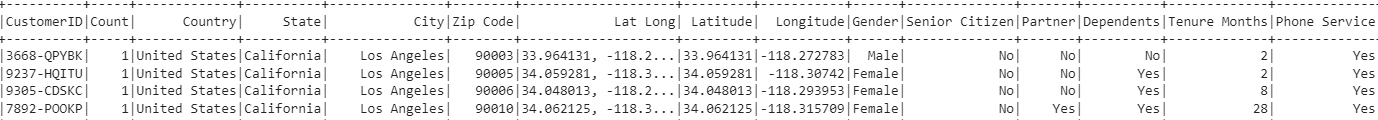
**II. Data Description**  
***Data Source:*** The analysis was conducted on the ‘Telco\_customer\_churn.csv’ and ‘Telco\_customer\_churn\_status.csv’ dataset which was sourced from IBM accelerator catalog - <https://accelerator.ca.analytics.ibm.com/bi/?perspective=authoring&pathRef=.public_folders%2FIBM%2BAccelerator%2BCatalog%2FContent%2FDAT00148&id=i9710CF25EF75468D95FFFC7D57D45204&objRef=i9710CF25EF75468D95FFFC7D57D45204&action=run&format=HTML&cmPropStr=%7B%22id%22%3A%22i9710CF25EF75468D95FFFC7D57D45204%22%2C%22type%22%3A%22reportView%22%2C%22defaultName%22%3A%22DAT00148%22%2C%22permissions%22%3A%5B%22execute%22%2C%22read%22%2C%22traverse%22%5D%7D>

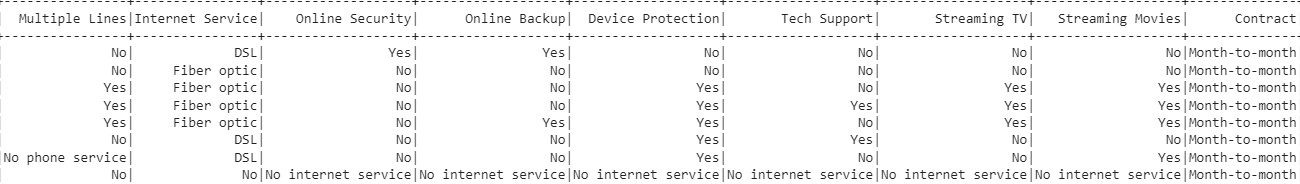
***Data Description:*** The customer churn dataset contains information about customers of a telecommunications company. The dataset includes various attributes related to customer demographics, services subscribed, billing information, and churn status. The dataset includes below variables -

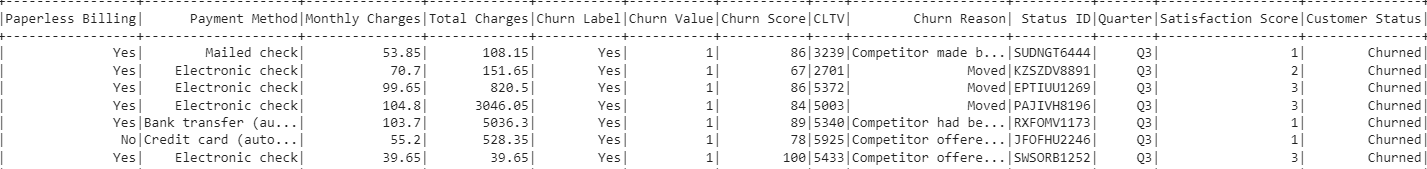
1. CustomerID: Unique identifier for each customer. String
2. Count: May represent the number of times a record appears in the dataset (usually 1 for unique rows). Numerical
3. Country: Country of the customer's residence. String
4. State: State of the customer's residence. String
5. City: City of the customer's residence. String
6. Zip Code: Postal code of the customer's residential address. Numerical
7. Lat Long: Combined latitude and longitude coordinates of the customer's location. Numerical
8. Latitude: Latitude component of the customer's location.Numerical
9. Longitude: Longitude component of the customer's location.Numerical
10. Gender: Gender of the customer (e.g., Male, Female, Non-Binary). Categorical
11. Senior Citizen: Indicates if the customer is considered a senior citizen (usually a binary indicator). Categorical
12. Partner: Indicates if the customer has a partner (e.g., Yes, No). Categorical
13. Dependents: Indicates if the customer has dependents (e.g., Yes, No).Categorical
14. Tenure Months: Number of months the customer has been with the company. Numerical
15. Phone Service: Indicates if the customer subscribes to phone service (e.g., Yes, No). Categorical
16. Multiple Lines: Indicates if the customer subscribes to multiple telephone lines (e.g., Yes, No, No phone service). Categorical
17. Internet Service: Type of internet service the customer subscribes to (e.g., DSL, Fiber optic, No). Categorical
18. Online Security: Indicates if the customer subscribes to an additional online security service (e.g., Yes, No, No internet service). Categorical
19. Online Backup: Indicates if the customer subscribes to an online backup service (e.g., Yes, No, No internet service). Categorical
20. Device Protection: Indicates if the customer subscribes to a device protection plan (e.g., Yes, No, No internet service). Categorical
21. Tech Support: Indicates if the customer subscribes to technical support services (e.g., Yes, No, No internet service). Categorical
22. Streaming TV: Indicates if the customer subscribes to streaming TV services (e.g., Yes, No, No internet service). Categorical
23. Streaming Movies: Indicates if the customer subscribes to streaming movies services (e.g., Yes, No, No internet service). Categorical
24. Contract: Type of customer contract (e.g., Month-to-month, One year, Two year). Categorical
25. Paperless Billing: Indicates if the customer has opted for paperless billing (e.g., Yes, No). Categorical
26. Payment Method: Method of payment (e.g., Electronic check, Mailed check, Bank transfer, Credit card). Categorical
27. Monthly Charges: The amount charged to the customer monthly. Numerical
28. Total Charges: The total amount charged to the customer over the life of the account. Numerical
29. Churn Label: Indicates if the customer has churned (e.g., Yes, No).Categorical
30. Churn Value: A numerical indicator of churn (e.g., 1 for churned, 0 for not churned). Categorical
31. Churn Score: A score representing the likelihood of churn. Numerical
32. CLTV: Customer Lifetime Value indicating the total worth to a business of the customer over the whole period of their relationship. Numerical
33. Churn Reason: The reason the customer churned (various categories). Categorical
34. Status ID: An identifier that could be linked to customer status in the company's internal system. Numerical
35. Quarter: The fiscal quarter in which the data was recorded or the customer status was updated. Categorical
36. Satisfaction Score: A score representing the customer's satisfaction with the service. Numerical
37. Customer Status: Current status of the customer (e.g., Active, Churned). Categorical
38. Churn Category: A categorization of churn reasons. Categorical
39. CustomerGroup: A classification of customers based on certain criteria (e.g., spending and tenure). Categorical
40. IsChurned: A binary indicator of whether the customer has churned (e.g., 1 for Yes, 0 for No). Categorical
41. TenureGroup: Grouping of customers based on tenure length. Categorical
42. MonthlyChargesRange: Categorization of monthly charges into ranges for analysis. Categorical
43. InvitationSource: The source from which the customer was invited or referred to the service. String
44. ReferralCustomerID: The customer ID of the person who referred the current customer, if applicable. String

***Sample Size and Variables:*** The dataset contains 7014 observations and 44 variables.

***Sample of Observations:***







***Interest in the Data:*** This dataset is crucial for understanding the dynamics of customer loyalty and designing strategies to improve retention. The customer churn dataset is of significant interest because it allows us to analyze factors that influence customer retention and churn in a telecommunications company. By understanding the patterns and characteristics of customers who churn, the company can take proactive measures to improve customer satisfaction, enhance loyalty, and reduce churn rates. This analysis can ultimately lead to better business decisions and increased profitability.

# **III. Research Questions**

1. Identify churn rates among different customer groups categorized based on their monthly charges and tenure
2. Is there a relationship between monthly charges and churn?
3. How does the source of invitation influence the churn rate among customers?
4. What are the predominant sources of referral?
5. Identifying customers who are likely to churn in the near future enables businesses to take proactive measures to retain them.

**IV. Methodology**

Our data analysis workflow was executed using PySpark, leveraging its distributed computing capabilities to handle large volumes of data efficiently. The following outlines our systematic approach:

**1. *Data Acquisition and Loading:*** We began by sourcing the data from IBM accelerator catalog, ensuring its integrity during the import process into a PySpark DataFrame.

**2. *Data Preprocessing:*** To prepare our data for analysis, we performed several preprocessing steps:

**i. *Data Transformation:*** Normalized data formats and transformed categorical variables into numerical format using one-hot encoding to facilitate mathematical computations in the modeling phase.

**ii. *Data Imputation***: Employed statistical methods, such as mean or median imputation, to fill in missing values, maintaining the integrity of our datasets.

**3. *Feature Engineering:*** Enhanced our model's predictive potential by:

**i. *Feature Selection:*** Applied selection techniques to identify the most significant predictors of churn, thereby optimizing our model's complexity and performance.

**4.** ***Exploratory Data Analysis (EDA):*** Conducted thorough exploratory analysis to uncover underlying structures and relationships within the data:

i. Utilized summary statistics and visualization tools within PySpark to extract insights and guide further analysis.

ii. Investigated correlations and performed hypothesis testing to assess the statistical significance of observed relationships.

iii. Model Development and Validation: Developed predictive models using machine learning algorithms available in the PySpark MLlib. We ensured the robustness of our models by:

iv. Splitting the data into training and test sets to evaluate performance and mitigate overfitting.

**6.** ***Interpretation of Results:*** Analyzed the output of our models to derive actionable insights. We interpreted the importance of different features and evaluated the model against key performance metrics.

Each step of our methodology was rigorously documented to ensure reproducibility and transparency in our analysis. The Python notebooks encompassing our full analytical process are available upon request for further examination.

**V. Results and Finding**

***1. Churn rates among different customer groups categorized based on their monthly charges and tenure:***

The "High Spend, Short Tenure" group has the highest churn rate, significantly more than any other group. This indicates that newer customers with higher monthly charges are more likely to leave the service.

***2. Relationship between monthly charges and churn:***

The churn rate increases with higher monthly charges, with the highest churn rate observed in the "Over 90" charges range. This indicates a possible price sensitivity among customers, where higher costs may lead to increased churn.

***3. Influence of source of invitation on churn rate among customers;***

The churn rate varies by invitation source, suggesting that how customers are referred to the service has an impact on their likelihood to churn.

"Merchant collaboration" is the referral source with the highest churn rate, while "friend of" has the lowest, indicating that customers referred through merchant collaborations might be less satisfied or have different expectations than those referred by friends.

***4. Predominant sources of referral:***

The most common source of referrals is from "family member of," suggesting that existing customers' family connections play a significant role in attracting new customers.

The referral source "referred by" has the lowest count, which could indicate that direct person-to-person referrals are less frequent compared to other methods like "promotional offer" or "merchant collaboration."

**VI. Conclusion**

1. The company should consider revising its pricing strategy for the highest-spending customer segments like giving some rewards.

2. The company should conduct a thorough evaluation of ‘merchant collaboration programs’ to identify factors contributing to higher churn rates and try mitigating them.

3. The company should incentivize "family member" referrals, which appear to be the leading referral source.

4. Predictive model can help identify customers who are likely to churn in the near future, businesses can take proactive measures to retain them.

**VII. References**

1. IBM #Cognos Analytics <https://community.ibm.com/community/user/businessanalytics/participate/blogs/blogviewer?BlogKey=FF811D76-ABE0-4DF2-BCEA-917176FD72E4>
2. What is Customer Churn Analysis https://www.gainsight.com/glossary/what-is-customer-churn-analysis/