Telco Customer Churn

IBM Exploratory Data Analysis for Machine Learning – Course Project

Hon Sing Sam Chung

07 March 2025

Dataset Overview & Data Source

Dataset Overview

- The Telco Customer Churn dataset contains over 7000 rows and 21 columns of data on telecommunications customers, focusing on predicting whether a customer will churn (leave the service).

Data Source

- Available on Kaggle, commonly used for churn prediction and customer retention analysis.

Key Attributes

CustomerID: Unique customer identifier.

Gender: Customer's gender.

SeniorCitizen: Whether the customer is a senior citizen.

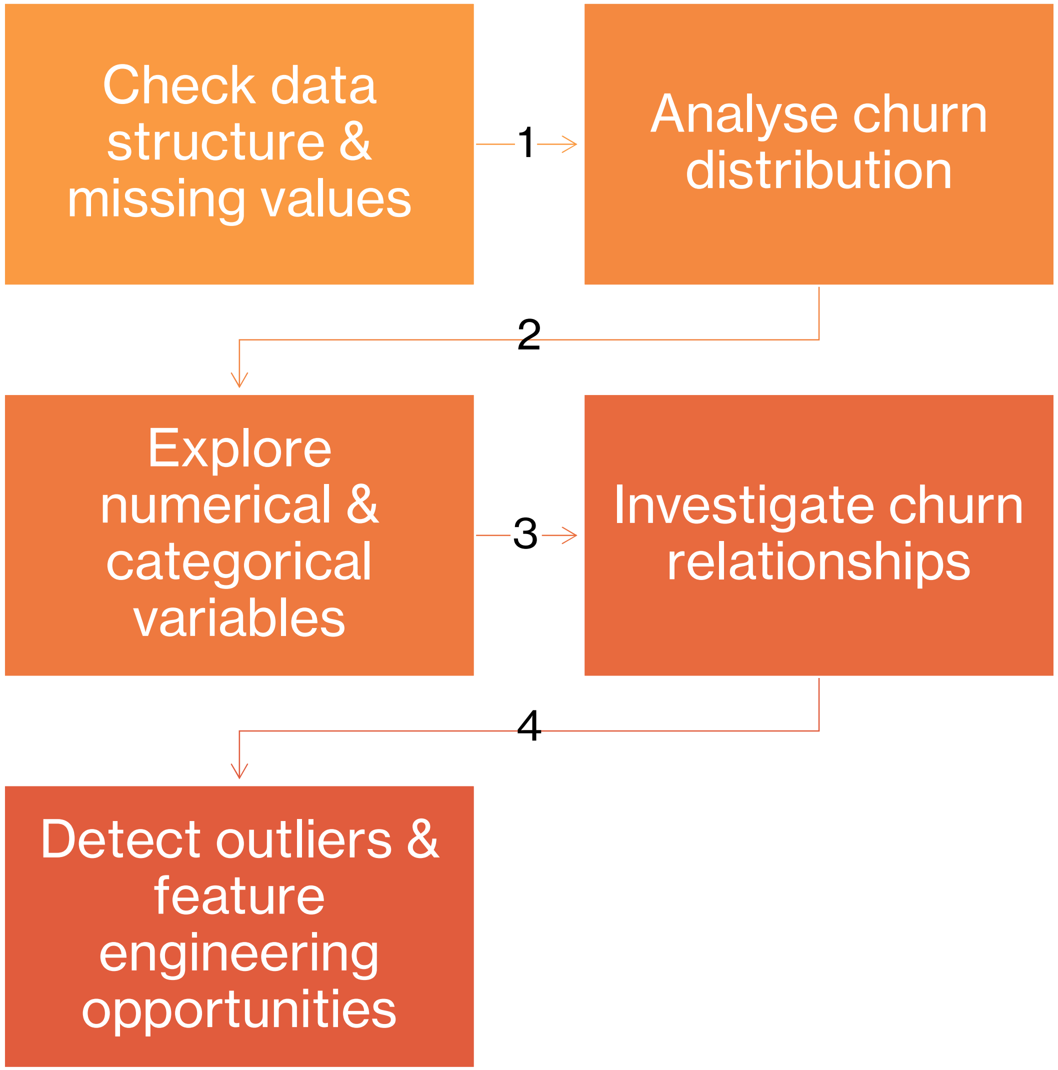
Tenure: Length of time as a customer (in months).

Contract: Type of customer contract.

Churn: Whether the customer has churned (target variable).

MonthlyCharges: The monthly charges for the customer (USD).

Exploratory Data Analysis (EDA) Plan



Data Cleaning & Preprocessing Steps

Handled missing values (e.g., imputed missing Total Charges)

Encoded categorical variables (one-hot encoding)

Removed duplicate records

Feature engineering

Feature Engineering

Tenure Groups

0-12 months = New

13-24 months = Established

25+ months = Long-Term

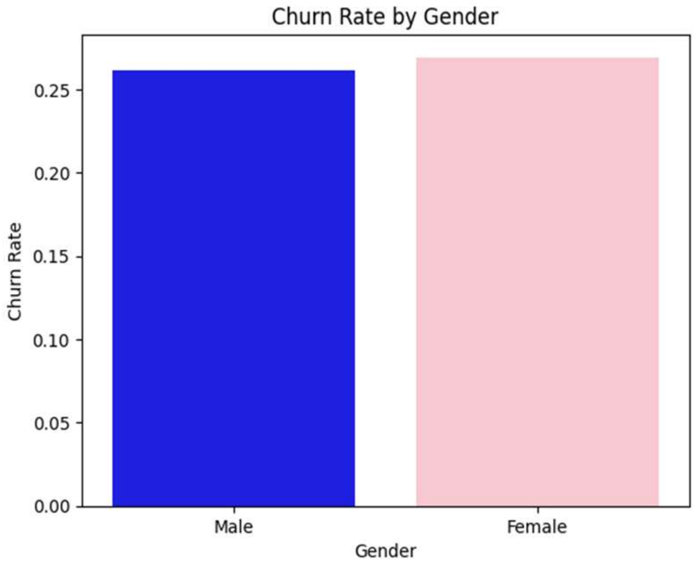
Charge Groups

Putting monthly charges into bins:

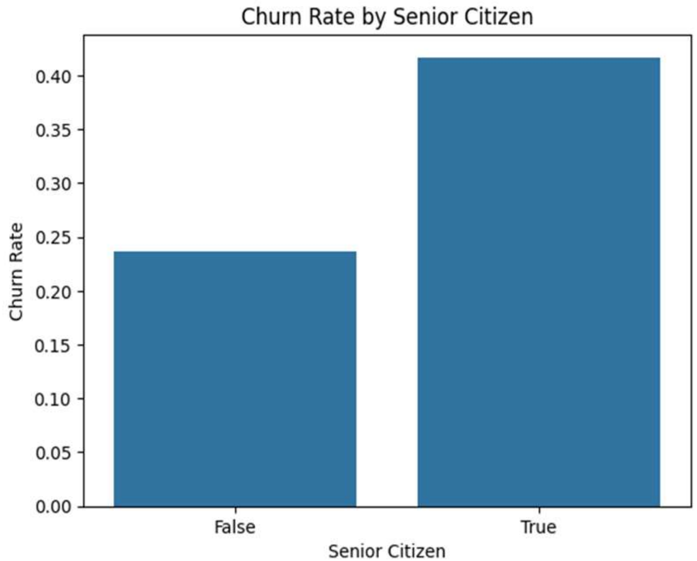
0-25, 26-50, 51-75, 76-100, 100+

Key Findings & Insights

Customer Demographics and Churn Patterns



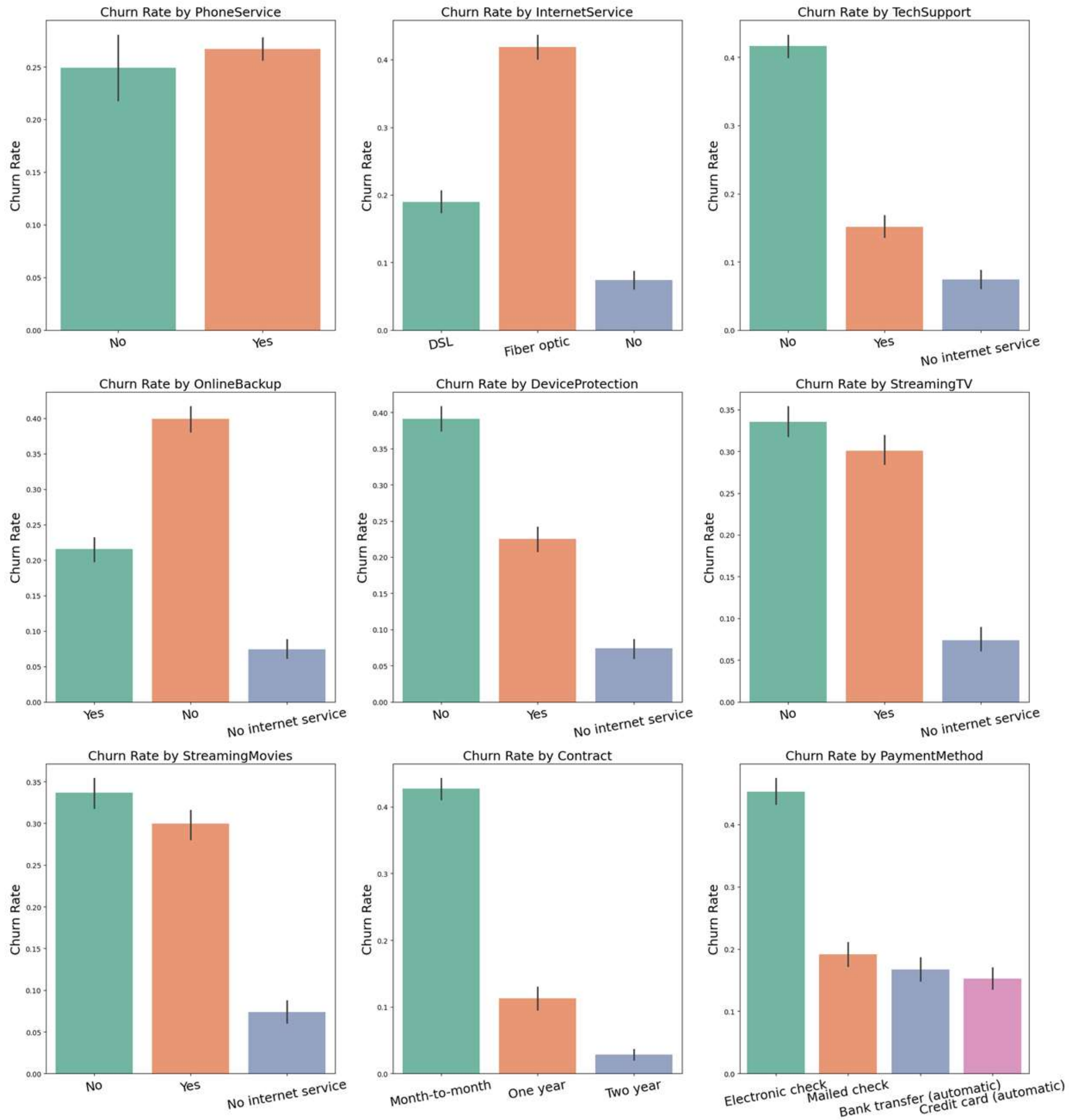
Gender does not appear to have a huge effect on churn rate



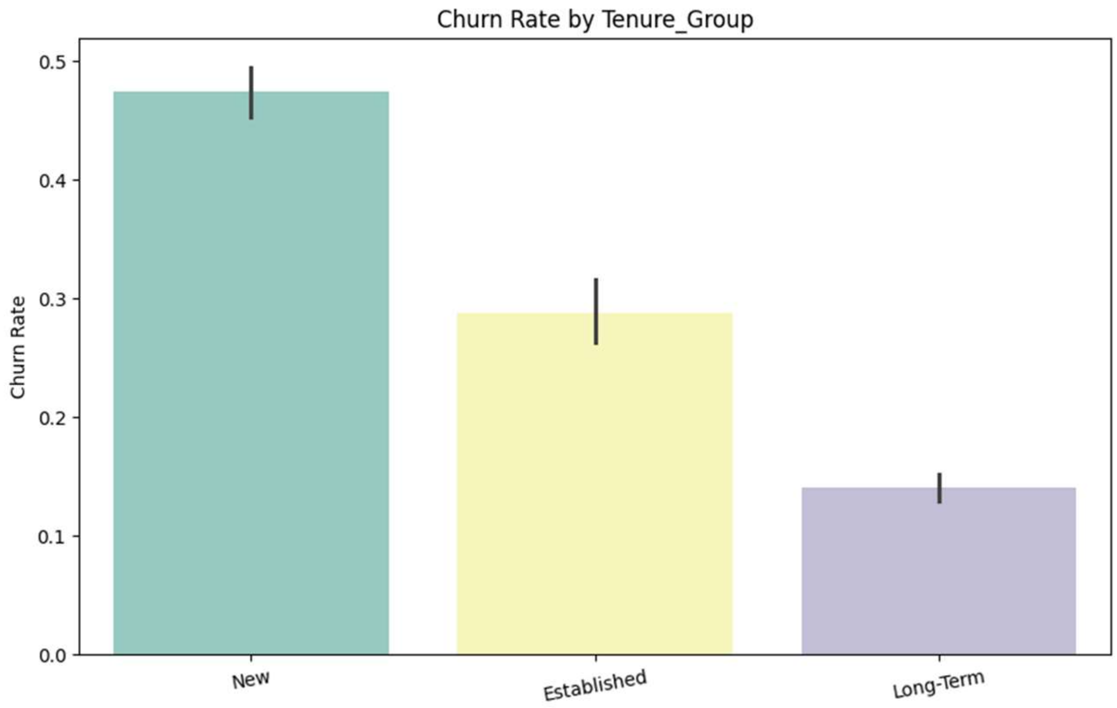
Senior Citizens tend to churn more often than non-Senior Citizens. This could suggest that more tailored services or retention strategies are needed for this group.

Churn by Service Features

- Large correlations between churn rate and those with Fiber Optic internet service, no Tech Support, No Device Protection, and No Online Backup.
    - Focus on improving customer engagement for customers with fewer services or offer discounts for bundling services together.
- The correlation between churn rate and month-to-month contracts is likely because a greater proportion of customers have month-to-month contracts. Since these customers are not bound by long-term commitments, they may be more likely to churn compared to those with longer-term contracts.
    - Encourage more customers to switch to 1-year or 2-year contracts, which may improve retention. This could also lead to more stable revenue.
- Those with no internet service appear to have the lowest churn rate.



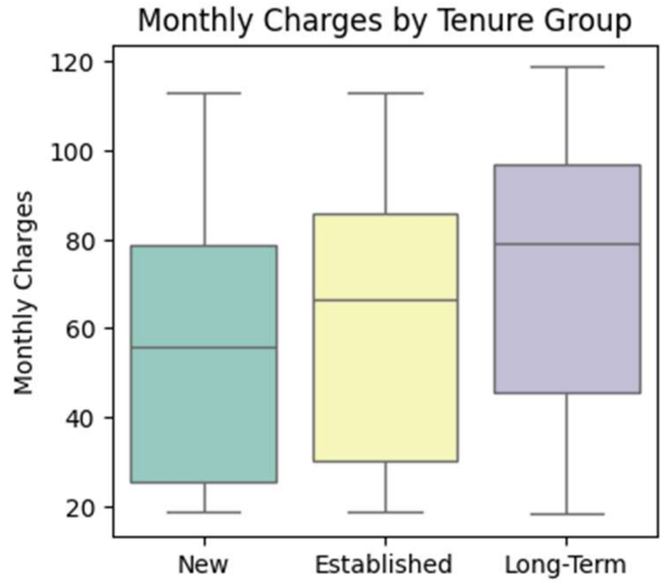
Tenure Insights

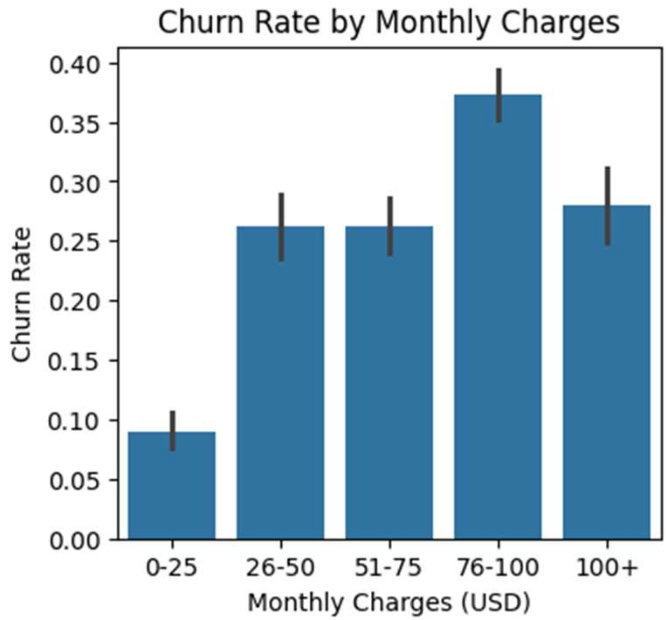


- Newer customers (low tenure) are more likely to churn than long-term customers.
- Focus on improving the customer experience and engagement within the first year to increase likelihood that customers will stay long-term.

Charges and Tenure

- Long-term customers (24+ months) have higher monthly charges compared to newer customers. Possibly due to upgrading services over time.
- New customers (0-12 months) tend to have lower monthly charges, possibly due to promotional pricing, introductory plans, or opting for basic or lower-cost plans when they first join.
- The lowest churn rate is observed in the 0-25 monthly charge range, which suggests that customers who pay lower amounts tend to stay longer, possibly due to being more satisfied with affordable plans.





Charges and Tenure (cont.)

- Focus on Retaining Low-Cost Customers:
  - Offer loyalty rewards, free upgrades, or discounts for long-term commitment.
  - Ensure exceptional customer service to maintain satisfaction.
- Review High-Cost Plans for Retention:
  - Investigate customer satisfaction with high-price plans through surveys.
  - Introduce flexible pricing or alternatives to prevent churn.
- Consider Bundling Services:
  - Offer bundled packages for high-charge customers to improve value perception and reduce churn.
- Promote Special Offers for New Customers:
  - Use discounts or promotional plans to retain new customers and prevent early churn.
- Create Targeted Retention Campaigns:
  - Focus retention efforts on high-churn customers in the high-price bracket.
- Develop a Customer Loyalty Program:
  - Offer exclusive perks or service tiers to long-term, high-paying customers to enhance loyalty.
- Test Pricing Strategies:
  - Implement A/B testing for different pricing models and track which reduces churn effectively.

Hypotheses

Hypothesis 1:	**H_0 (Null Hypothesis):** Senior citizens do not have a higher churn rate than non-senior citizens. (There is no significant difference in churn rates between senior and non-senior customers.) **H_1 (Alternative Hypothesis):** Senior citizens do have a higher churn rate than non-senior citizens.
Hypothesis 2:	**H_0 (Null Hypothesis):** The churn rate of customers with Fiber Optic internet is not significantly different from the churn rate of customers with DSL or No Service. **H_1 (Alternative Hypothesis):** The churn rate of customers with Fiber Optic internet is significantly higher than that of customers with DSL or No Service.
Hypothesis 3:	**H_0 (Null Hypothesis):** Customers with month-to-month contracts do not churn at a higher rate than those with 1-year or 2-year contracts. **H_1 (Alternative Hypothesis):** Customers with month-to-month contracts churn at a higher rate than those with 1-year or 2-year contracts.

Statistical Analysis of Hypothesis 3

Performed the Chi-Square test for Contract Type vs Churn.

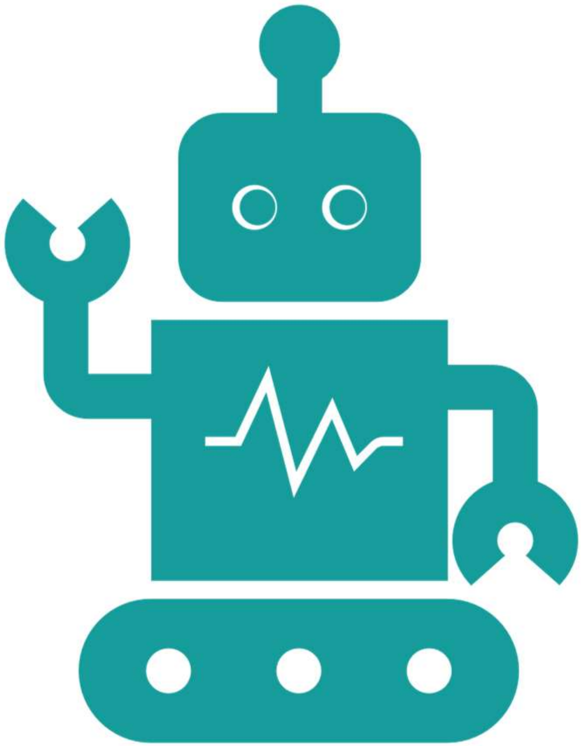
- Chi-Square Statistic of 1184.6 shows a strong relationship between Contract Type and Churn.
- Extremely low p-value of 5.86e-258 further supports the observed differences are not due to chance.
- We reject the null hypothesis and conclude that customers with month-to-month contracts churn at a higher rate than those with 1-year or 2-year contracts.

```
contingency_table = pd.crosstab(df['Contract'], df['Churn'])
chi2_stat, p_value, dof, expected = chi2_contingency(contingency_table)
print("Chi-Square Statistic:", chi2_stat)
print("P-value:", p_value)
print("Degrees of Freedom:", dof)
print("Expected Frequencies:\n", expected)
```

```
Chi-Square Statistic: 1184.5965720837926
P-value: 5.863038300673391e-258
Degrees of Freedom: 2
Expected Frequencies:
 [[2846.69175067 1028.30824933]
 [1082.11018032  390.88981968]
 [1245.198069    449.801931  ]]
```

Next steps for further analysis & improvements

- Train machine learning models
  - Build models (Logistic Regression, Random Forest) to predict churn.
  - Perform cross-validation to assess model performance and prevent overfitting.
- Improve feature engineering
  - Combine related services into bundled features for more meaningful insights.
- Conduct A/B testing on retention strategies
  - Such as discounts or service bundling, to assess their impact on reducing churn.



Summary

The Telco Customer Churn dataset is relatively clean and contains a variety of valuable features, making it a solid foundation for churn analysis. Missing values have been appropriately handled, and categorical variables have been encoded. However, some features, such as "TotalCharges," required imputation where no charges were recorded as ‘ ‘ (empty space, not null, not 0), and there may be additional data on customer interactions, satisfaction, or specific service usage that could further enhance the model's accuracy. Access to more granular data, like customer support interactions or detailed usage patterns, would help refine retention strategies and improve predictive model performance.