BENTLEY UNIVERSITY

Churn Insights

Unveiling customer behavior to drive retention and profitability.

Sahil Bora & Vinit Late

MSBA Candidates’ 2025 | Bentley University

Churn

Churn refers to the rate at which customers stop doing business with a company over a specific period. It reflects customer attrition and is commonly used as a metric to measure customer retention. High churn rates indicate that many customers are leaving, which can impact revenue and growth, making it a critical concern for businesses in competitive markets like telecommunications.

Why does it matter ?

Revenue Impact: Losing customers directly affects profitability.

Competitive Market: Customers have numerous options and can easily switch to competitors.

Retention Advantage: Identifying and targeting “high-risk” customers early can optimize retention strategies and reduce costs.

Business Challenge - Telecom Industry

Industry Churn Insight: Annual churn rate in the telecom industry ranges from 15–25%, posing a significant challenge in retaining customers

Retention vs. Acquisition: Retaining existing customers is far more cost-effective than acquiring new ones, yet requires targeted strategies.

How can we predict customer churn, identify underlying customer behavior patterns, and segment customers to develop targeted retention and profitability strategies?

What factors significantly influence Monthly Charges?

What are the key predictors of customer churn ?

What insights can be gained to optimize revenue across customer groups?

Are demographics key factors ?

Can statistics can shape the business?

Can the dataset be reduced to fewer dimensions while retaining key patterns and variability

What meaningful labels can be assigned to clusters, and how do they inform marketing and service strategies?

Understanding the Data

This study utilizes a telecommunications dataset consisting of 7,043 observations across 21 variables, which include customer demographics, account details, service usage patterns, and churn status

The data set includes information about:

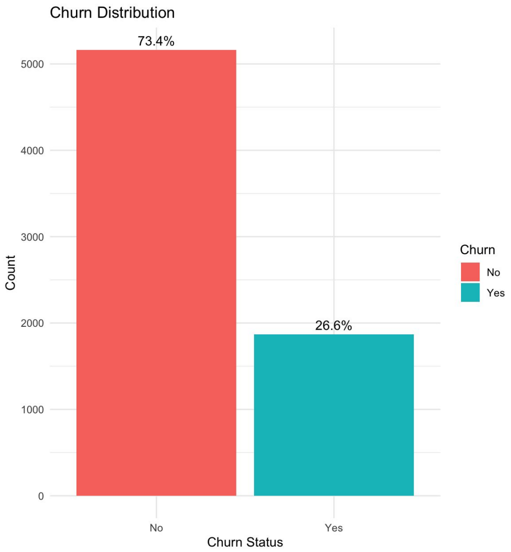
- **Customers who left within the last month** – the column is called Churn
- **Services that each customer has signed up for** – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- **Customer account information** - how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- **Demographic info about customers** – gender, age range, and if they have partners and dependents

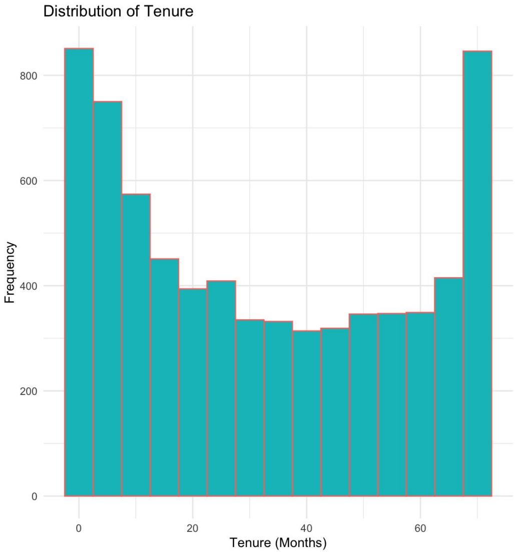
Data Handling

- **Handle Missing Values (Direct and Indirect)**
- **Data Manipulation**

Exploratory Data Analysis

Understanding the context



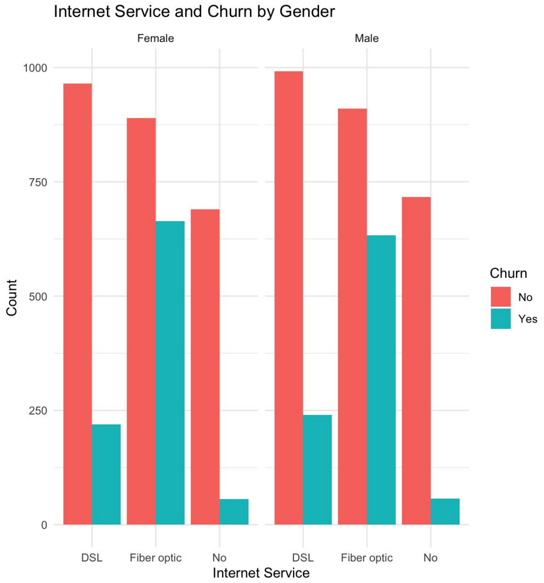


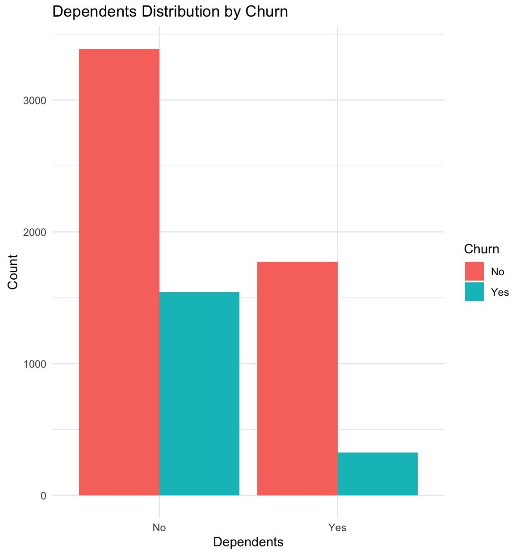
Approximately 73.4% of customers have not churned ("No"), while 26.6% have churned ("Yes"). This highlights a significant proportion of retained customers but also indicates potential revenue risks from the churned segment

The graph shows a higher concentration of customers with very short tenure (around 0 months) and very long tenure (close to 72 months), with fewer customers in the mid-range tenure. This indicates that most customers either churn quickly or stay loyal for extended periods, suggesting the need for tailored strategies for mid-tenure engagement.

Exploratory Data Analysis

Understanding Churn



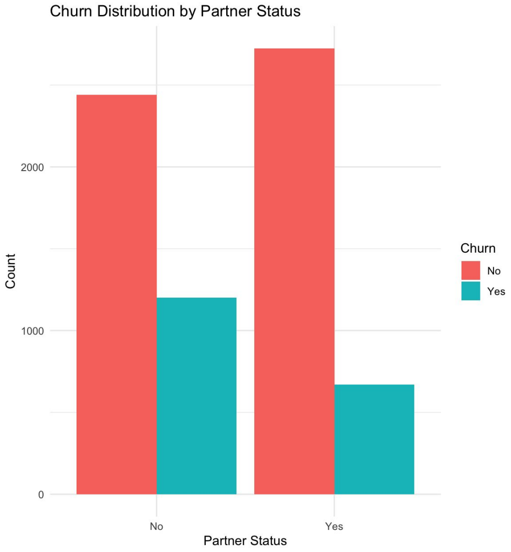


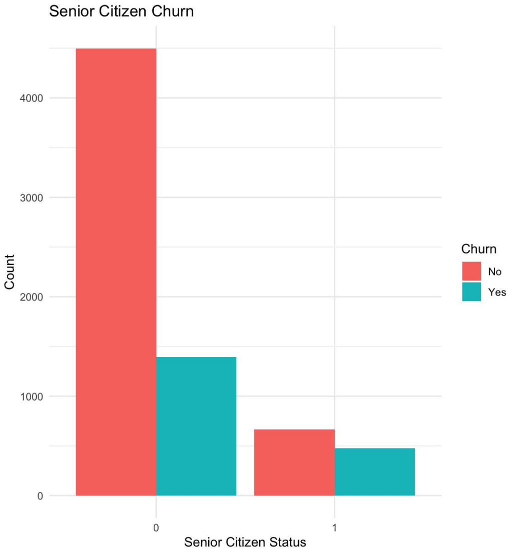
This bar chart shows the distribution of churn segmented by gender and type of internet service. Customers with "Fiber Optic" service have a noticeably higher churn rate compared to "DSL" or "No Internet Service." This indicates that the churn problem is more pronounced among Fiber Optic users, suggesting potential issues with service satisfaction or cost for this group.

Customers without dependents show a higher churn rate compared to those with dependents. This suggests that having dependents may contribute to customer stability, possibly due to bundled services or shared usage among family members

Exploratory Data Analysis

Understanding Churn



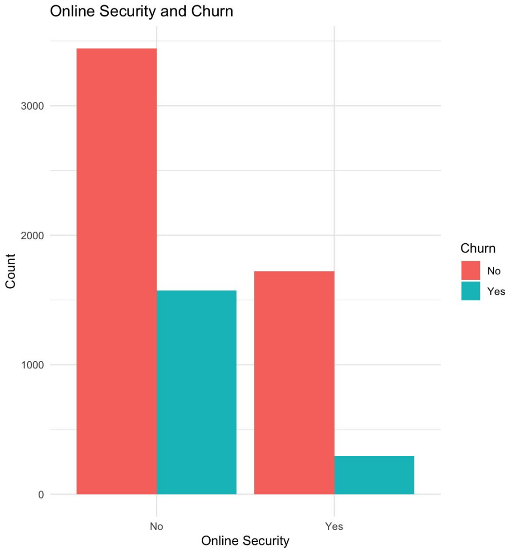


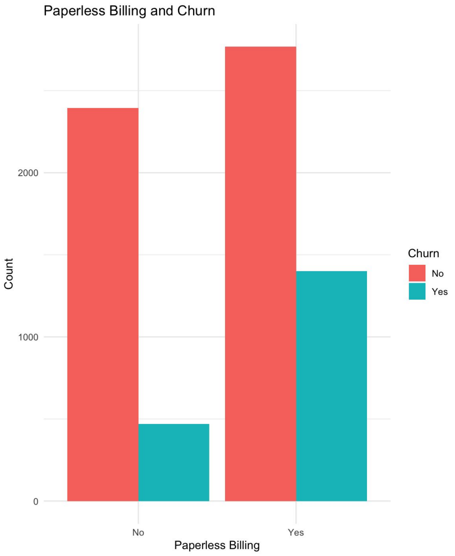
Customers without partners have a higher churn rate compared to those with partners. This indicates that having a partner may contribute to customer retention, possibly due to shared usage of services or a stronger perceived value of the services.

While the majority of customers are non-senior citizens and show lower churn rates, senior citizens have a relatively higher proportion of churn. This suggests that senior citizens might face barriers or challenges that increase their likelihood of leaving, such as pricing or service usability

Exploratory Data Analysis

Understanding Churn



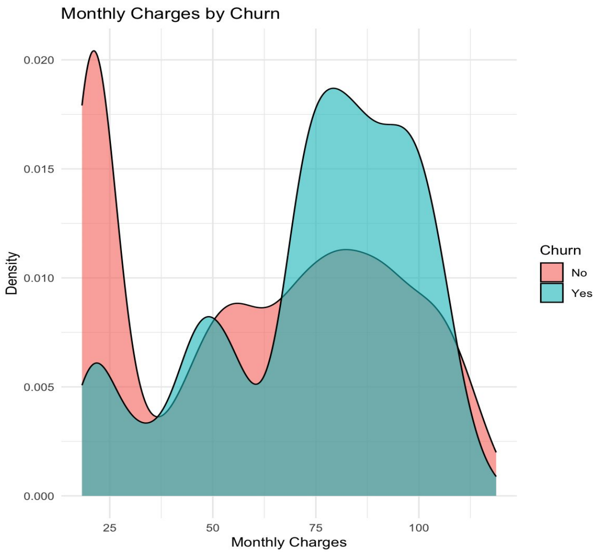


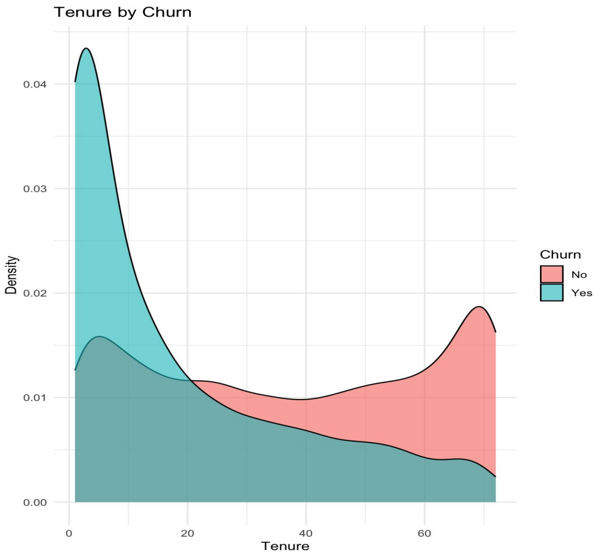
Customers without online security exhibit significantly higher churn rates compared to those with online security. Providing or improving online security services could be a key strategy to reduce churn and retain customers.

Customers who use paperless billing have a significantly higher churn rate compared to those who do not. This suggests that addressing issues related to paperless billing, such as ease of use or communication, may help reduce churn.

Exploratory Data Analysis

Understanding Churn





Customers with higher monthly charges have a greater tendency to churn. Churned customers are concentrated around mid-to-high charge ranges, while non-churned customers are distributed more broadly, including lower monthly charges. This suggests that pricing and perceived value at higher charge levels may be significant factors influencing churn

Customers with shorter tenures are more likely to churn, as the density for churned customers is higher at lower tenure values. In contrast, customers with longer tenures tend to stay, as the density for non-churned customers increases with tenure length. This suggests that loyalty programs or retention strategies should focus on newer customers.

Statistical Modelling

This analysis employed a range of statistical and machine learning techniques to explore the drivers of customer churn and develop actionable customer segments. Each method was chosen to address specific aspects of the data and analysis objectives.

Multiple Linear Regression is used to predict MonthlyCharges based on factors such as tenure, InternetService, Contract, etc. This approach helped identify key drivers of billing patterns and provided insight into how different variables influence customer charges.

To understand the factors affecting customer churn, a **Logistic Regression** model was applied. This model evaluated the likelihood of churn using variables such as tenure, MonthlyCharges, InternetService, Contract, Dependents etc. The results from this model pinpointed significant predictors and allowed for probability-based classification of customer behavior.

Principal Component Analysis (PCA) was performed to reduce the dimensionality of numerical variables, including tenure, MonthlyCharges, and TotalCharges. By extracting key components, PCA revealed the variables most responsible for explaining variance in the data, enabling better visualization and understanding of underlying patterns.

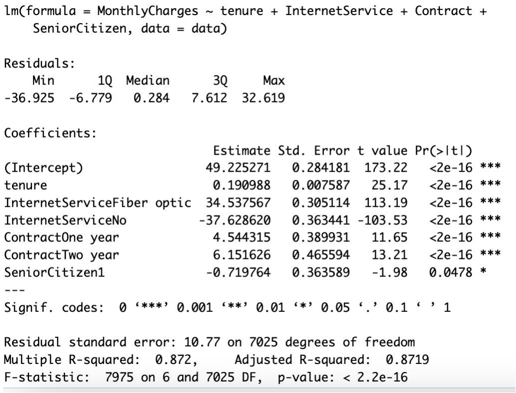
Finally, **K-Means Clustering** was used to group customers into distinct segments based on tenure, MonthlyCharges, and TotalCharges. This unsupervised learning approach highlighted three primary customer groups, offering valuable insights for targeted marketing and retention strategies. Together, these methods provided a comprehensive framework for understanding customer behavior and informing actionable strategies.

Multiple Linear Regression

What factors significantly influence Monthly Charges?

- Tenure: Positive coefficient (0.191), meaning that customers with longer tenure tend to have slightly higher charges.
- Customers with Fiber Optic services pay significantly more ($34.54 higher on average).
- Customers without Internet Service pay significantly less (-$37.63 on average).
- One-year contracts result in an average increase of $4.54 in monthly charges.
- Two-year contracts result in an average increase of $6.15.

Key Takeaway: Internet service type is the most influential predictor of monthly charges, followed by tenure and contract type.

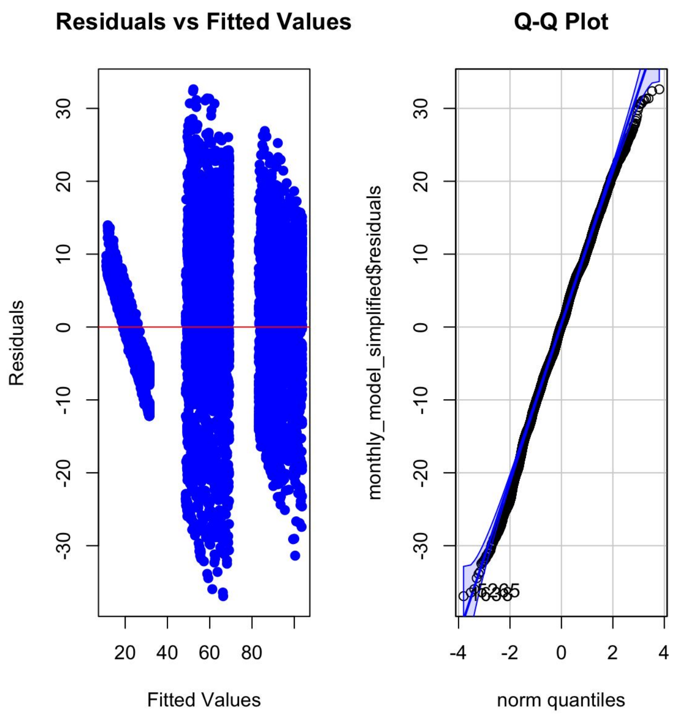


```
lm(formula = MonthlyCharges ~ tenure + InternetService + Contract +
    SeniorCitizen, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-36.925  -6.779   0.284   7.612  32.619

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 49.225271   0.284181  173.22   <2e-16 ***
tenure                       0.190988   0.007587   25.17   <2e-16 ***
InternetServiceFiber optic  34.537567   0.305114  113.19   <2e-16 ***
InternetServiceNo          -37.628620   0.363441 -103.53   <2e-16 ***
ContractOne year             4.544315   0.389931   11.65   <2e-16 ***
ContractTwo year             6.151626   0.465594   13.21   <2e-16 ***
SeniorCitizen1              -0.719764   0.363589   -1.98   0.0478 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.77 on 7025 degrees of freedom
Multiple R-squared:  0.872,     Adjusted R-squared:  0.8719
F-statistic:  7975 on 6 and 7025 DF,  p-value: < 2.2e-16
```

Model Interpretation

- The model predicts MonthlyCharges using five predictors: tenure, Internet Service, Contract, and Senior Citizen.
- The Adjusted **R-squared value is 0.872**, indicating that 87.2% of the variability in MonthlyCharges is explained by the model. This shows a strong fit to the data.
- The **Residual Standard Error is 10.77**, suggesting moderate variance in unexplained data.
- All predictors have **VIF values below 2.5**, indicating low multicollinearity among variables.

Multiple Linear Regression


Residuals vs Fitted Values

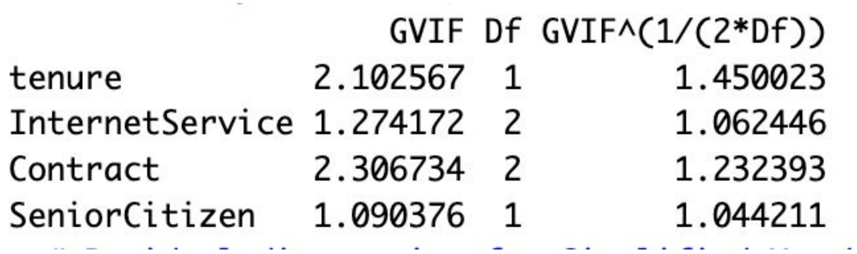
Q-Q Plot

The residuals are randomly scattered around the horizontal line at 0, suggesting that the model captures the linear relationship between predictors and the target variable well. However, some variability at extreme fitted values

The residuals mostly align with the diagonal line, indicating that they are approximately normally distributed. Slight deviations at the tails suggest potential outliers or non-normality in extreme cases

The model assumptions of linearity and normality are largely satisfied, making the model reliable for inference and predictions

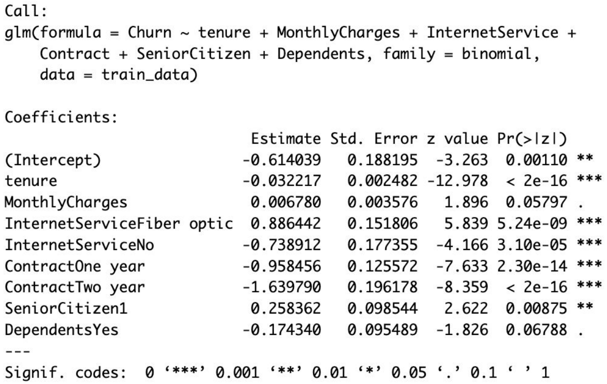
VIF Values


```
                    GVIF Df GVIF^(1/(2*Df))
tenure          2.102567  1        1.450023
InternetService 1.274172  2        1.062446
Contract        2.306734  2        1.232393
SeniorCitizen   1.090376  1        1.044211
```

Logistic Regression

What are the key predictors of customer churn:

- Tenure: Longer tenure reduces the likelihood of churn (odds ratio: 0.97), indicating that longer relationships increase customer loyalty.
- Internet Service: Customers with fiber optic internet are more likely to churn (odds ratio: 2.43), while those without internet service are less likely to churn (odds ratio: 0.48).
- Contract Type: Longer contracts significantly reduce churn. One-year contracts have an odds ratio of 0.38, and two-year contracts are even lower at 0.19.
- Senior Citizen Status: Senior citizens are slightly more likely to churn (odds ratio: 1.29).
- Monthly Charges: Higher charges slightly increase churn likelihood (odds ratio: 1.006), but the effect is marginal.

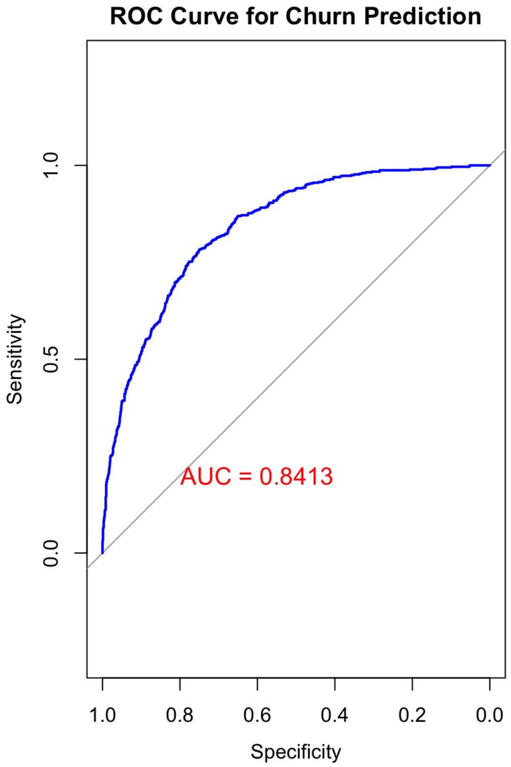


```
Call:
glm(formula = Churn ~ tenure + MonthlyCharges + InternetService +
    Contract + SeniorCitizen + Dependents, family = binomial,
    data = train_data)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.614039   0.188195  -3.263  0.00110 **
tenure                     -0.032217   0.002482 -12.978  < 2e-16 ***
MonthlyCharges              0.006780   0.003576   1.896  0.05797 .
InternetServiceFiber optic  0.886442   0.151806   5.839 5.24e-09 ***
InternetServiceNo          -0.738912   0.177355  -4.166 3.10e-05 ***
ContractOne year           -0.958456   0.125572  -7.633 2.30e-14 ***
ContractTwo year           -1.639790   0.196178  -8.359  < 2e-16 ***
SeniorCitizen1              0.258362   0.098544   2.622  0.00875 **
DependentsYes              -0.174340   0.095489  -1.826  0.06788 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model Interpretation

- The model achieves an accuracy of 79.79% and an AUC of 0.84, indicating good predictive performance.
- High sensitivity (90.05%) suggests that the model effectively identifies non-churning customers, though specificity (51.43%) is lower, meaning some churners are missed
- The confusion matrix reveals that high-risk customers are characterized by short tenure, fiber optic internet, and month-to-month contracts

Logistic Regression



The Area Under the Curve (AUC) is 0.8413, indicating a strong performance of the logistic regression model.

The curve is above the diagonal line (random classifier), which confirms that the model performs better than random guessing

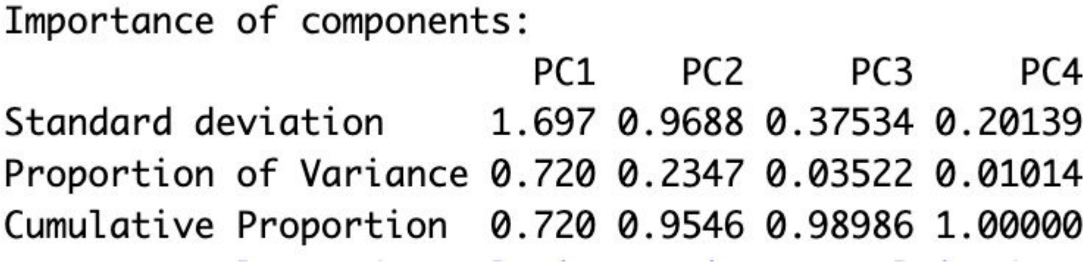
Principal Component Analysis

Can the dataset be simplified to fewer dimensions?

- Yes, the dataset can be effectively simplified to two principal components (PC1 and PC2) since these components explain a significant portion of the total variance (95.46%). Dimensionality reduction helps in simplifying the dataset while retaining most of its information.

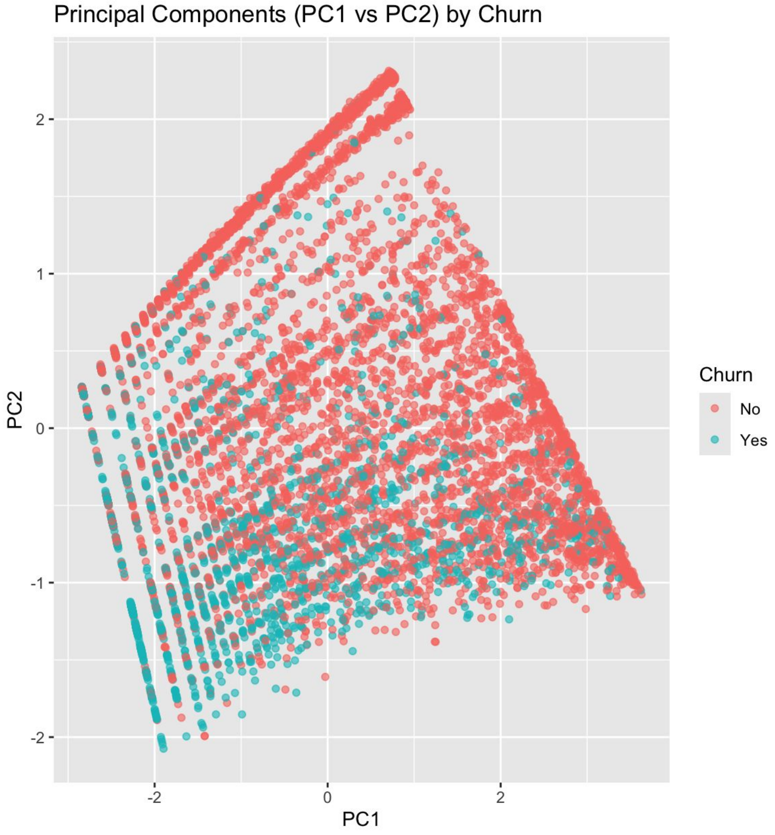
What proportion of variance is explained by each principal component?

- **PC1: Explains 72.0% of the variance**, which indicates it captures most of the variability in the dataset.
- **PC2: Explains an additional 23.47%** of the variance.
- Cumulative Variance (PC1 + PC2): Together, PC1 and PC2 explain 95.46% of the total variance, making them highly representative of the dataset.



```
Importance of components:
                         PC1    PC2     PC3     PC4
Standard deviation     1.697 0.9688 0.37534 0.20139
Proportion of Variance 0.720 0.2347 0.03522 0.01014
Cumulative Proportion  0.720 0.9546 0.98986 1.00000
```

Principal Component Analysis



This indicates that churn and no-churn customers share significant similarities in the primary dimensions (PC1 and PC2)

Non-churned customers ("No") are more densely packed and spread across the upper-right quadrant, indicating stability in certain regions of the principal component space.

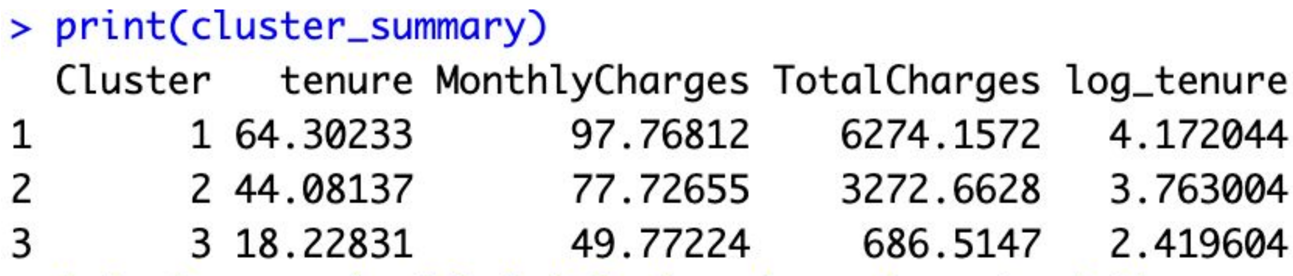
Churned customers ("Yes") appear more scattered and concentrated in specific areas, hinting at potential customer segments prone to churn.

K-means Clustering

Can customers be grouped based on behavior?

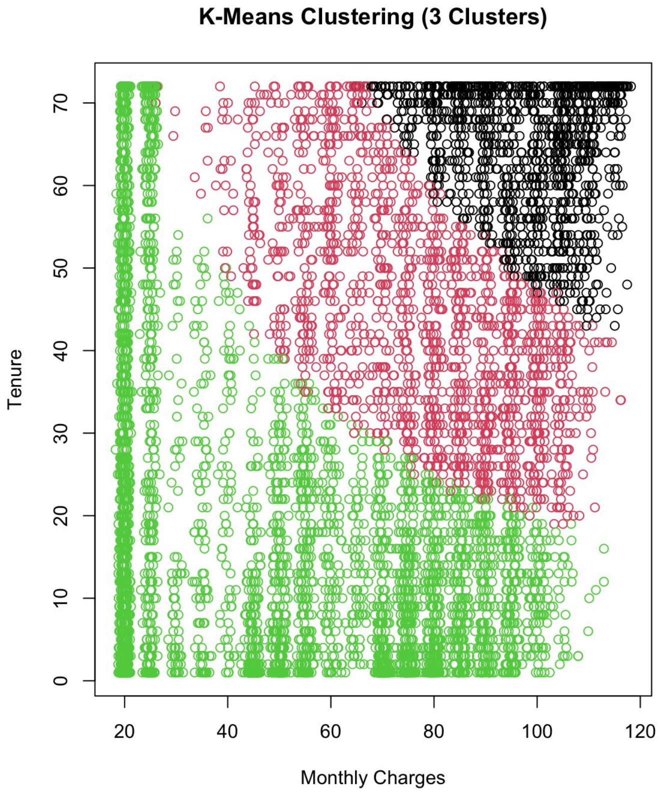
Customers were grouped based on tenure, MonthlyCharges, and TotalCharges (numeric attributes). Removing outliers ensured that extreme values did not skew the clustering. The three clusters were identified as:

- **High Spenders (Cluster 1):** Customers with high tenure, high MonthlyCharges, and high TotalCharges.
- **Discount Seekers (Cluster 2):** Customers with moderate values for tenure, MonthlyCharges, and TotalCharges.
- **New Customers (Cluster 3)**: Customers with the lowest tenure and relatively low spending.



```
> print(cluster_summary)
  Cluster   tenure MonthlyCharges TotalCharges log_tenure
1       1 64.30233       97.76812    6274.1572   4.172044
2       2 44.08137       77.72655    3272.6628   3.763004
3       3 18.22831       49.77224     686.5147   2.419604
```

K means Clustering



The relationship between Monthly Charges and Total Charges is linear within clusters, as expected.

- Cluster 1 (High Spenders) spans higher ranges for all variables.
- Cluster 3 (New Customers) is concentrated in the lower range of tenure and charges.

Improvements

Data Enhancements:

- Collect additional customer data, such as age or feedback scores, to uncover more behavioral patterns and improve the accuracy of our analyses.

Variable Refinements:

- Break down Total Charges into more granular metrics (e.g., monthly average spending or frequency of upgrades).

Cluster Analysis Refinement:

- Revisit the clustering process to evaluate whether merging or splitting clusters based on business needs (e.g., merging similar segments like Discount Seekers and New Customers) could provide better actionable insights.

Model Fine-Tuning:

- Apply hyperparameter tuning for the logistic regression model (e.g., threshold optimization) to improve sensitivity and specificity.

Future Scope

Real-Time Analytics:

- Implement a streaming pipeline to monitor customer behavior in real-time, enhancing the ability to intervene before churn occurs.
- Integrate a live dashboard for visualizing key metrics like churn probability and customer cluster movement.

Customer Behavior Tracking:

- Develop longitudinal studies on cluster transitions (e.g., tracking New Customers' journey toward becoming High Spenders).
- Analyze customer feedback for sentiment trends to incorporate qualitative data into predictive models.

Custom Retention Strategies:

- Create segment-specific marketing campaigns targeting High Spenders, Discount Seekers, and New Customers.

Cross-Validation of Findings:

- Validate the current findings using additional datasets or cross-validation techniques to ensure generalizability.
- Expand the dataset to include more granular features like service complaints or geographic location for refined clustering.

Closing

Key Takeaways:

- The analyses reveal that tenure, contract type, and internet service are critical predictors of churn.
- PCA successfully reduced dimensionality while preserving ~95% variance, providing actionable insights for visualization and interpretation.
- K-Means clustering grouped customers into three meaningful segments, supporting tailored retention strategies.

Business Implications:

- Retention efforts should prioritize High Spenders (Cluster 1), while personalized engagement is crucial for New Customers (Cluster 3).
- Insights from regression models and clustering empower targeted marketing and service upgrades.

In conclusion, this comprehensive analysis highlights the critical drivers of customer churn and provides actionable strategies for targeted retention and engagement. By leveraging regression, PCA, and clustering, we've uncovered key behavioral insights that align with business goals. Moving forward, integrating these findings into data-driven decision-making processes will help improve customer satisfaction, reduce churn rates, and drive sustainable growth.

Sahil
Bora

Vinit
Late

THANK YOU