**BCIS 5110 Programming Languages for Business Analytics**

**Final Project Report**

**“Customer Churn Prediction for Telco”**

**Project by**:

Sai Teja Kode

**Table of Contents:**

1. Executive summary………………………………………………………………..3
2. Project motivation/background……………………………………………………3
3. Data description…………………………………………………………………....3
4. Data preparation …………………………………………………………………..5
5. Exploratory data analysis………………………………………………………….5
6. Models and analysis………………………………………………………………12
7. Findings and managerial implications……………………………………………19
8. Conclusion………………………………………………………………………..20
9. Appendix…………………………………………………………………………22
10. References………………………………………………………………………..22
11. **Executive Summary**

The following project is on analyzing and predicting customer churn for a telecommunications company using the Telco Customer Churn dataset. The main aim is to finding out the most important factors or predictors of the reason for the customer churn by exploring customer demographics, subscription services, and account information. Additionally, predictive models are developed to assist in identifying targeted retention strategies. Actionable insights derived from the analysis will support informed managerial decision-making..

1. **Project Motivation/Background**

Customer churn is definitely one of the major problems faced my any company which effects their profitability. Prediction and prevention of customer churn can be one of the crucial business priorities. This project has been designed to answer two basic questions:

1. What factors are most influential in customer churn?
2. How accurately can we predict customer churn using the available data?

By understanding customer behaviors and implementing predictive modeling, this project aims at identifying high-risk customers, thus allowing for precise and effective retention strategies.

1. **Data description:**

The Telco Customer Churn dataset consists of 7,043 rows (observations) and 21 columns (variables). The primary focus is on predicting customer churn, denoted by the Churn variable (Yes/No). Key features include:

1. Categorical Variables These variables capture customer demographics and account-related factors, such as:
   1. Demographics: Gender, SeniorCitizen, Partner, and Dependents.
   2. Services: InternetService, Contract, and PaymentMethod.
   3. Outcome: Churn (target variable).
2. Numerical Variables These variables represent customer engagement and financial metrics:
   1. Tenure: Duration or the time (in months) for which the customer has been with the company.
   2. MonthlyCharges: Monthly billing amount.
   3. TotalCharges: Total billing amount since joining.
3. Unique Identifier
   1. customerID: A unique identifier assigned to each customer.

By examining these features, the analysis aims to determine relationships between customer characteristics, their service usage patterns, and their propensity to churn.

**Datatypes and their description**:

1. customerID: String & Categorical – Primary key or the Unique identifier for each customer
2. gender: String & Categorical – Represents Gender of each customer which can be either male or female.
3. SeniorCitizen: Integer & Categorical – whether the customer is a senior citizen or not. Values are either 0s(No) or 1s(Yes)
4. Partner: String & Categorical – shows whether a customer has a partner or not. Values are either No or Yes.
5. Dependents: String & Categorical - shoes Whether the customer has any dependents. Values are either No or Yes.
6. tenure: Integer & Numerical - Tenure a customer has stayed with the company in months.
7. PhoneService: String & Categorical – Shows whether a customer is equipped with phone service or not. Values are either No or Yes.
8. MultipleLines: String & Categorical - Shows if a customer has multiple lines Values are either No or Yes.
9. InternetService: String & Categorical – service type of the internet.
10. OnlineSecurity: String & Categorical – shows if a customer service plan has online security or not
11. OnlineBackup: String & Categorical - shows if a customer service plan has Online Backup or not
12. DeviceProtection: String & Categorical - shows if a customer service plan has Device Protection or not.
13. TechSupport: String & Categorical - shows if a customer service plan has Tech Support or not.
14. StreamingTV: String & Categorical - shows if a customer has Streaming TV or not.
15. StreamingMovies: String & Categorical - shows whether a customer has Streaming Movies or not.
16. Contract: String & Categorical – the type of the contract based on the time frame.
17. PaperlessBilling: String & Categorical – Shows whether the customer opted for paperless billing or not.
18. PaymentMethod: String & Categorical – shows the type of payment method opted by the customer.
19. MonthlyCharges: Float & Numerical – Charges incurred by the customer monthly.
20. TotalCharges: Float & Numerical – total charges incurred by the customer
21. Churn: String & Categorical – The dependent variable shows whether a customer has churned or not.
22. **Data Preparation**

Below-mentioned steps are taken to clean and prepare the data:

1. Handling Missing Values:  
   Missing values or the blank values in the TotalCharges column were replaced with the mean values using fillna() function.
2. Standardizing Inconsistent Values:  
   There are inconsistencies in relation to a “no internet service” value present in the few columns or variables such as MultipleLines, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies. The above-mentioned inconsistencies related to the “no internet service” values are handled by replacing the value with “no.” In other words, treating the “no internet service” values the same as the value “no.”
3. Type Conversion: Converted values in the TotalCharges column to a numeric data type. Also encoded categorical variables using label encoding method.
4. Feature Engineering: Added a new “tenure\_group” variable to group customers based on their tenure intervals of [0-12 months, 12-24 months, 24-48 months, and 48-72 months] by assigning bins and labels . Also converted SeniorCitizen and other features to categorical or numerical types as required.
5. Service Aggregation:  
   Separately created a TotalServices column which represents the total number of services that a customer subscribed to. Generated a ChurnBinary column for binary classification (1 for churned, 0 otherwise).
6. Scaling: Applied standard scaling to numerical variables (MonthlyCharges, TotalCharges) for model input consistency.
7. **Exploratory Data Analysis: Statistics and Graphs:**
8. Overview of Churn Distribution:

* Around 26%(1869) of customers in the dataset had churned (`Churn = Yes`), while the remaining 74%(5163) were retained (`Churn = No`).
* The imbalance in the churn distribution highlights the need for evaluation metrics like AUC-ROC and F1-Score in predictive modeling.

Graph: Below Bar chart was created to visualize the distribution of churned and non-churned customers.

A graph with a red and blue rectangle

Description automatically generated

1. Tenure Analysis:

* As per the below-plotted boxplot, the average tenure of churned customers was turned out to be 17.98 months, which is significantly lower than the 37.65 months for non-churned customers.
* Lower average tenure for the churned customers represents that customers who are with the shorter tenures are more likely to churn.

Graph: Below boxplot has been plotted for comparing the tenure of churned vs. non-churned customers that highlights the disparity in tenure.

A diagram of a chart

Description automatically generated with medium confidence

1. the distribution of customers based on their tenure with the company:

* New Customers: This is the largest group, with over 2000 customers, and probably consists of those customers who have joined recently. This means that quite a significant number of the customer base comprises newbies.
* 1-2 Years: A fair drop from that of new customers, yet it still retains a remarkable count, showing that some retention of customers has taken place for a year or two.
* 2-3 Years and 3-4 Years: These ranges are trending downwards, indicating fewer customers stay with the service for this length of tenure.
* 4-5 Years: A slight increase from the previous categories, which postulates that some customers tend to stick with the service after a few years.
* Loyal Customers: This category also has a considerable number of customers,whose tenure is more than 4-5 Years, highlighting a very strong base of long-term loyal customers.

Graph: Below bar chart represents the distribution of customers across different tenure ranges, categorizing customers based on their tenure, or time as a customer, and the count of customers falling into each category..

A graph of customer distribution across tenue range

Description automatically generated

1. Monthly Charges and Total Charges:

The customers with higher `MonthlyCharges` were more likely to churn in other words higher the monthly charges more likely the customers to churn and `TotalCharges` showed a positive correlation with longer-tenured customers.

Average total charges varied across different contract types:

* Month-to-Month: The average total charges are lowest for customers on a month-to-month contract, $1,369.25. This may indicate lesser commitment on the part of these customers due to flexibility or shorter duration of service.
* One Year: The customers on a one-year contract have an average of much higher total charges at $3,034.68, higher than those on month-to-month contracts. These types of contracts promise a longer commitment, and hence higher charges are accumulated.
* Two Year: The highest average total charges are for two-year contracts, which is $3,728.93. This indicates that the longer the contract, the more revenue per customer, as one would expect.

Graph: below bar plot has been generated to compare the average total charges across different contract types.

A graph of a chart

Description automatically generated with medium confidence

1. Distribution of customers across different contract types:

* The highest churn rate was among customers who had "Month-to-Month" contracts.
* One-Year and Two-Year contract customers had significantly fewer churn rates.

Graph: This bar chart shows the Customer Distribution Across Contract Types, reflecting the number of customers categorized by their contract type.

A chart with green bars

Description automatically generated

1. Correlation between the number of services customers subscribe to and their likelihood of churning:

* A weak negative correlation of -0.07 was found between the total number of services subscribed and churn, indicating that customers subscribing to more services were slightly less likely to churn.
* As per the below bar chart, Higher the number of services subscribed, lower the churn rate.

- Graph: Below bar chart of churn rates by the number of subscribed services was created to see the relationship visually.

A graph of a number of services

Description automatically generated

1. Customer Segmentation Based on Service Type:

* Phone Only: This category accounts for a fair number but stands considerably lesser than "Internet+Add-ons."
* Internet Only: The smallest customer base, showing that stand-alone internet services are less in demand compared to bundled services.
* Internet + Add-ons: There is the highest number of subscribers (approx. 4000). The people are thus oriented towards internet services paired with added advantages.

Graph: the following bar diagram represents the Segregation of customers based on the type of service.

A graph of a customer service

Description automatically generated

1. **Models and Analysis: Predictions and Accuracy Evaluation**

1. **Data Preparation for Modeling**

Before building the predictive models, some essential preprocessing steps were done for the quality and model readiness of the data:

Feature Encoding: The categorical variables like gender, Partner, and Contract have been transformed into numerical values using the label encoding method to prepare them for the machine learning algorithms.

Feature Selection: The important predictors that were chosen based on their relevance to the customer churn included tenure, MonthlyCharges, TotalCharges, and Contract.

Train-Test Split: The dataset was split into two subsets—70% for training the model and 30% for testing—to evaluate its performance on unseen data.

Feature Scaling: Numerical variables (MonthlyCharges, TotalCharges) were standardized to ensure uniformity in their scale, preventing features with larger magnitudes from dominating the model.

---

**2**. **Models Built**

1. **Logistic Regression:**

ROC Curve:

A ROC curve shows the graphical performance of a model against various thresholds for classification. The y-axis on the ROC curve depicts the True Positive Rate or sensitivity, and the x-axis depicts the False Positive Rate. The curve shows a good separation or the distinction between the positive and negative classes.

A graph with a line

Description automatically generated

AUC-ROC Score:

The AUC is 0.84, indicating strong discriminatory ability. An AUC of 0.84 means that a model can correctly identify a positive class from a negative class 84% of the time.

**2. Classification Metrics**

A screenshot of a computer

Description automatically generated

1. Precision:
   * the ratio of (true positives) in the numerator to the sum of (true positives and false positives) in the denominator.
   * For class 0 (non-churn): Precision = 0.84, meaning 84% of the non-churn instances predicted were correct.
   * For class 1 (churn): Precision = 0.65, meaning 65% of the instances predicted as churn were correct.
2. Recall (Sensitivity):
   * Recall calculates a ratio of actual positives that were predicted correctly.
   * For Class 0: Recall = 0.89 - Non-churn instances were 89% correctly identified.
   * For Class 1: Recall = 0.53 - Churn instances were correctly identified at 53%. This may imply poor performance in terms of recall of true instances of churn.
3. F1-score:
   * The F1-score = harmonic mean of (precision and recall).
   * For class 0: F1-score = 0.87, showing strong performance for non-churn cases.
   * For class 1: F1-score = 0.59, reflecting relatively lower performance due to lower recall.
4. Confusion Matrix:
   * The confusion matrix summarizes the model's predictions:
   * True Negatives (1386): Correctly predicted non-churn.
   * False Positives (163): Incorrectly predicted churn for non-churn customers.
   * False Negatives (261): Missed churn customers.
   * True Positives (300): Correctly predicted churn customers.
   * While the model performs well for non-churn (class 0), it struggles to accurately predict churn (class 1).

**Random Forest Classifier**:

ROC Curve:

A ROC curve shows the graphical performance of a model against various thresholds for classification. In the ROC curve, the y-axis is the True Positive Rate, which gives the sensitivity, while the x-axis is the False Positive Rate. In y axis, true positive rate is shown and in x axis false positive rate is shown . The curve shows a good separation or the distinction between the positive and negative classes.

A graph with a line

Description automatically generated

AUC-ROC Score:

The AUC is 0.82, indicating a good performance. An AUC = 0.82 means that the 82% of the time the current model correctly identifies variation between the positive and negative classes.

2. Classification Metrics

A screenshot of a computer screen

Description automatically generated

1. Precision:
   * Precision represents the ratio of predicted positives(numerator) that are actual positives.
   * For class 0 (non-churn): Precision = 0.83, meaning 83% of the predicted non-churn instances are correct.
   * For Class 1(Churn): Precision = 0.62; it means that 62% of the predicted instances for a churn are correct.
2. Recall Sensitivity: True positives /(false negatives +true positives).
   * For class 0: Recall = 0.89, which is very strong for its class and shows strength in catching the non-churn cases.
   * Class 1: Recall = 0.49: the model detects only 49% of actual cases of churn, which, for an application needing to detect churn, might not be good enough.
3. F1-Score: .
   * For Class 0: F1-score = 0.86, robust for its class.
   * For Class 1: F1-score = 0.55: This is a decent result for churn detection because of the lower recall value.
4. Confusion Matrix
   * True Negatives (1379): Correctly predicted non-churn cases.
   * False Positives (170): Predicted to churn for non-churn customers.
   * False Negatives (287): Missed the actual churn customers.
   * True Positives (274): Correctly predicted churn customers.

The model performs well on non-churn customers but could not identify many of the actual churn cases.

**Feature Importance - the most significant predictors of customer churning:**

A screenshot of a computer

Description automatically generated A graph with blue bars

Description automatically generated

The graph depicts the most significant predictors of customer churning. MonthlyCharges and TotalCharges are the most influential, with 19.41% and 19.03%, respectively, which means financial factors drive churn to a great extent. Tenure, at 16.48%, also plays a critical role in that longer customer relationships reduce churn risk. Contract type (9.37%) strongly impacts retention, especially for month-to-month customers

c. **Comparison Between Models:**

The Logistic Regression model produced a score of 80% accuracy and an AUC-ROC score of 0.84, demonstrating good discrimination quality. It yielded a higher recall for non-churn (0.89) but failed to achieve churn with a lower recall of 0.53. Compared to the Random Forest Classifier, which achieved lower metrics at 78% accuracy and AUC-ROC score of 0.82, but has the overall best feature interpretability. Random Forest performed at a similar recall for the non-churn class (0.89) but the individual prediction of churn resulted in retaining a recall of 0.49, which is slightly worse than that of Logistic Regression. Yet, in more complex relationships, Random Forest had value feature importance, which explained the predictor importance among MonthlyCharges, TotalCharges, making it a better option where detail and explainable insights are required. Conversely, Logistic Regression scored better in simplicity and predictive accuracy.

1. **Findings and Managerial Implications:**

**Key Findings**:

* Customer Tenure and Churn: Customers with shorter tenures (average ~18 months) are more likely to churn, whereas long-tenured customers (average ~38 months) exhibit stronger loyalty. Retaining new customers is a critical focus area to mitigate churn.
* Contract Type and Churn: Month-on-month basis contracts show the highest churn rates. Customers with One-Year or Two-Year contracts churn less frequently due to long-term commitments.
* Monthly Charges and Churn: High monthly charges are strongly correlated with churn. Customers paying premium rates without bundled services are at higher risk of leaving.
* Service Subscriptions and Churn: Bundled service subscribers (e.g., Internet + Add-ons) are less prone to churn. Customers with minimal services like "Phone Only" or "Internet Only" face a higher risk of leaving.
* Predictive Features: Key predictors of churn include `MonthlyCharges`, `TotalCharges`, `tenure`, `Contract`, and `InternetService`. These actionable features support tailored retention strategies.
* Correlation Analysis: A weak negative correlation (-0.07) exists between the number of subscribed services and churn. While bundling services marginally reduces churn, its impact isn't strongly significant.
* Churn Distribution by Customer Type: "New Customers" (tenure < 12 months) represent a disproportionate share of churn. Early engagement is essential for retaining these customers.
* Feature Importance: Random Forest analysis reveals that variables like `MonthlyCharges` and `Contract` significantly influence churn. These insights guide retention-focused decision-making.

**Managerial Implications:**

* Customer Retention Strategies: Prioritize retention of new customers by introducing early-stage incentives and personalized strategies, especially for customers with high `MonthlyCharges` on Month-to-Month contracts.
* Contractual Improvements: Promote One-Year and Two-Year contracts through attractive discounts and additional benefits. Offer flexible upgrade options for customers on Month-to-Month plans.
* Service Bundling: Emphasize the value of bundled services (e.g., Internet + Add-ons) to enhance customer loyalty. Highlight cost-effectiveness and added value in marketing efforts.
* Pricing Adjustments: Reassess pricing models for high-risk customers, particularly those with high `MonthlyCharges` and Month-to-Month contracts. Introduce loyalty rewards or value-added services that will improve retention.
* Proactive Customer Engagement: Use predictive analytics to identify and address high-risk customers. Utilize customer feedback in resolving pain points and making improvements proactively.
* Data-Driven Decision Making: Apply insights from feature importance analysis to refine marketing strategies and customer support. Focus on key retention drivers, such as competitive pricing and bundled services.
* Churn Prevention Campaigns: Launch tailored campaigns targeting high-risk customer segments like "Phone Only" and "Internet Only" users. Emphasize the advantages of upgrading to comprehensive service bundles in these campaigns.

1. **Conclusion**:

The Telco Customer Churn project successfully identified critical factors influencing customer churn and developed predictive models to assist in proactive retention strategies. Through an extensive analysis of the dataset, it was evident that variables such as contract type, monthly charges, tenure, and service subscriptions significantly impacted churn rates. Customers with Month-to-Month contracts and high monthly charges, particularly those with shorter tenures, were found to be at the highest risk of churning. These insights underscore the importance of targeting these customer segments with tailored offers and incentives to reduce churn effectively.

The project employed Logistic Regression and Random Forest models, achieving high predictive accuracy (~78%) and identifying key predictors of churn. Random Forest, in particular, provided valuable insights into feature importance, highlighting the influence of `MonthlyCharges`, `TotalCharges`, and `tenure` on churn behavior. The statistical and graphical explorations further enriched the understanding of customer behaviors, offering actionable insights for customer segmentation and service enhancement.

Despite challenges such as imbalanced data and the limited scope of variables, the project laid a strong foundation for implementing data-driven strategies. By focusing on high-risk customer segments and promoting bundled services and long-term contracts, telecom companies can significantly improve retention and profitability. Moving forward, incorporating additional customer behavior data and exploring advanced machine learning models could further refine the predictive capabilities and enhance real-time churn monitoring. Overall, this project demonstrates the immense potential of data analytics in addressing critical business challenges and driving customer loyalty.

1. **Appendix**:
2. **Dataset**: Telco Customer Churn <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>



1. Python file(.ipynb):   
   
2. **Sources**:
3. <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>
4. <https://unt.instructure.com/courses/108446/modules>
5. https://www.hackerearth.com/practice/machine-learning/machine-learningalgorithms/tutorial-random-forest-parameter-tuning-r/tutorial/