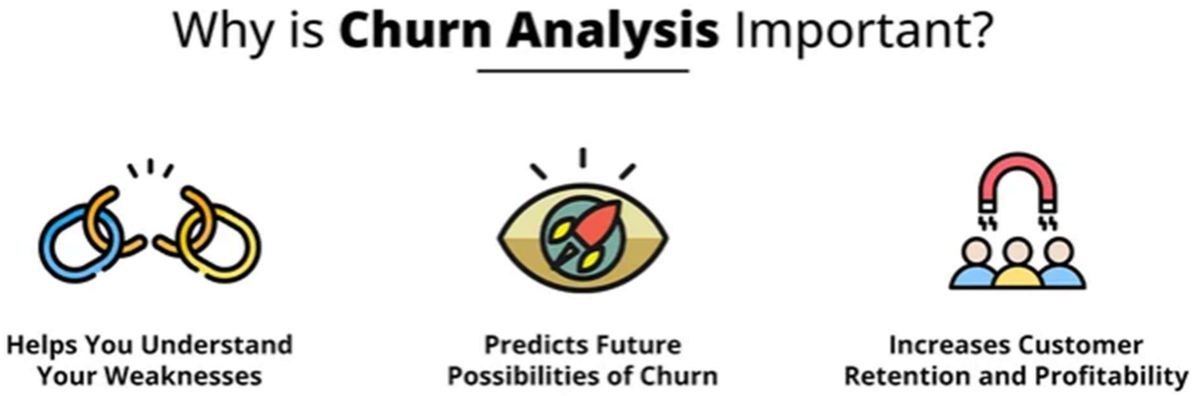
Customer Retention Strategies: A Comprehensive Churn Analysis

(Utilizing Machine Learning to Predict and Mitigate Churn)


Why is Churn Analysis Important?
Helps You Understand Your Weaknesses
Predicts Future Possibilities of Churn
Increases Customer Retention and Profitability

Ahmad Zalkat

EC Utbildning

Projekt i Data Science

2024-10

Abstract

This project focuses on analyzing customer churn data, aiming to extract insights into customer demographics, behavior, and contract patterns that lead to churn.

The dataset contains 6,418 entries, including customer details such as age, gender, contract type, and churn status.

The main objective is to identify the distribution of churn based on key factors like gender, payment method, contract type, and customer status.

The analysis also investigates the correlation between churn rate and these factors to derive actionable insights that could help reduce customer attrition

Förkortningar och Begrepp

- Churn: Loss of customers over time, crucial in telecommunications.
- Data Preprocessing: Cleaning and preparing data for analysis, including handling missing values.
- Machine Learning: AI branch for identifying patterns and making predictions based on data.
- Predictive Modeling: Forecasting future outcomes using statistical methods and algorithms.
- Confusion Matrix: A tool to evaluate model performance by showing correct and incorrect predictions.
- Precision: The accuracy of positive predictions made by the model.
- Overfitting: When a model performs well on training data but poorly on new data.
- Label Encoding: Converting categorical data into numerical format.
- Train-Test Split: Dividing data into training and testing sets to assess model performance.
- null: Indicates missing or absent data in a dataset.
- Total_Churn: The total number of customers who have churned (left the service).
- Churn_Rate: The percentage of customers who have churned, calculated as the number of churned customers divided by the total number of customers.
- Payment_Method: Refers to how customers make payments, like through bank transfers, credit cards, or paper checks.
- Contract: The type of contract a customer has, like month-to-month, one-year, or two-year agreements.
- Customer Status: The current state of a customer (e.g., Churned, Joined, Stayed).

Contents

Abstract 2

Förkortningar och Begrepp 3

1 Inledning 1

2 Teori 1

2.1 Churn Theory and Factors 1

2.1.1 Predictive Models for Churn 2

2.2 Metrics for Churn Evaluation 4

3 Metod 5

3.1 Data Collection 5

3.2 Data Preprocessing 5

3.3 Exploratory Data Analysis (EDA) 5

3.4 Churn Rate Analysis 6

3.5 Modeling Approach 6

3.6 Model Evaluation and Selection 7

3.7 Implementation of Results 7

4 Resultat och Diskussion 7

4.1 Results from Exploratory Data Analysis (EDA) 7

4.1.1 Key Findings: 7

4.1.2 Analysis: 7

4.2 Results from the Modeling Phase 8

4.2.1 Random Forest: 8

4.2.2 Logistic Regression: 8

4.2.3 Decision Trees: 8

4.2.4 Support Vector Machine (SVM): 9

4.2.5 Neural Networks: 9

4.3 Discussion 9

4.4 Power BI Visualization 10

5 Slutsatser 10

6 Självutvärdering 12

Appendix A 13

Källförteckning 15

1 Inledning

In this project, we aim to analyze customer churn data to understand the key factors that lead to customer retention and attrition. Churn, or customer attrition, refers to the process by which customers stop using a company's product or service. Understanding why customers leave is crucial for businesses, as it directly impacts revenue and growth.

The data set used in this analysis contains detailed information about customers, their demographics, service usage, payment methods, and reasons for churn. By examining these variables, we hope to uncover trends and patterns that will enable businesses to predict and mitigate churn, ultimately improving customer satisfaction and loyalty.

The analysis will involve data cleaning, exploratory data analysis (EDA), and the application of machine learning models to predict customer churn. In addition to quantitative insights, the project will provide visual representations of the data to better understand the relationships between various factors and churn rates.

This report is structured as follows:

- **Data Overview**: A summary of the dataset and its features.
- **Data Preprocessing**: Steps taken to clean and prepare the data for analysis.
- **Exploratory Data Analysis**: Insights drawn from initial examination of the data.
- **Modeling**: Predictive models used to identify factors contributing to churn.
- **Conclusion**: Summary of findings and recommendations for reducing churn.

2 Teori

Customer churn is a critical issue for many businesses, particularly those operating in competitive industries such as telecommunications, banking, and subscription-based services. Churn directly impacts a company's profitability, as acquiring new customers is often more costly than retaining existing ones. Therefore, reducing customer churn is a key focus for organizations seeking to improve their bottom line and sustain long-term growth.

The concept of churn is deeply rooted in customer relationship management (CRM), which focuses on understanding customer behavior and creating strategies to maintain customer loyalty. By identifying early signs of dissatisfaction or disengagement, businesses can intervene before a customer decides to leave. This process involves both qualitative and quantitative analysis of customer data to uncover factors that contribute to churn.

2.1 Churn Theory and Factors

Churn can be influenced by a variety of factors, including:

- **Customer Demographics:** Age, gender, marital status, and income level may all impact customer loyalty.

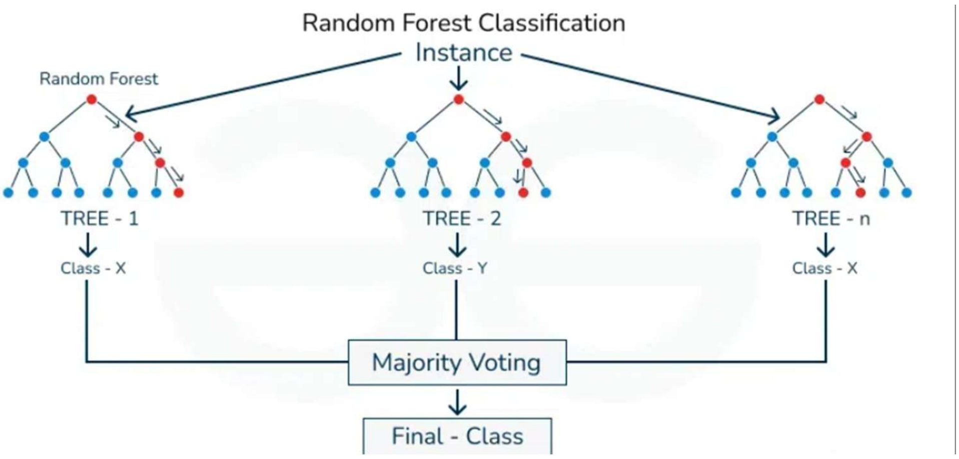
- **Service Usage:** The frequency and nature of service usage (e.g., internet services, premium support) often correlate with churn. Customers who are not fully engaged with a service or underutilizing it may be more likely to churn.
- **Pricing and Contracts:** Higher monthly charges, unfavorable contract terms, or lack of flexible payment options can push customers toward cancellation.
- **Customer Satisfaction**: Low satisfaction levels, often influenced by poor customer service or unresolved issues, are significant drivers of churn.
- **External Market Factors:** Competitive offers from other service providers can entice customers to switch.

2.1.1 Predictive Models for Churn

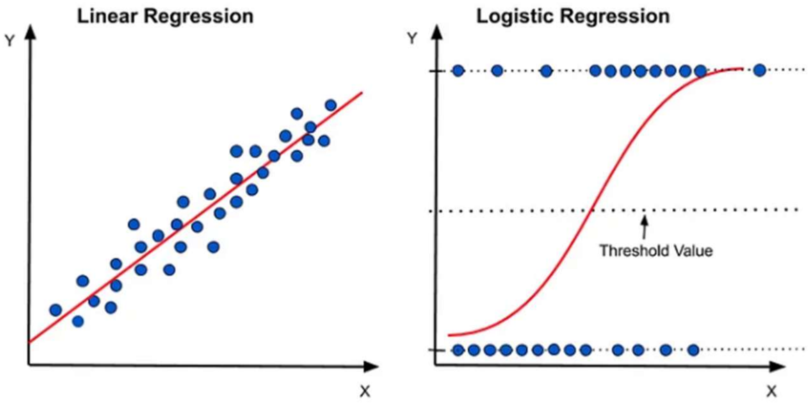
Theories behind predictive modeling suggest that customer churn can be predicted using various machine learning techniques. By analyzing historical customer data, we can develop models that assign a probability of churn to each customer. This allows businesses to target at-risk customers with retention strategies before they leave.

Common machine learning models used for churn prediction include:

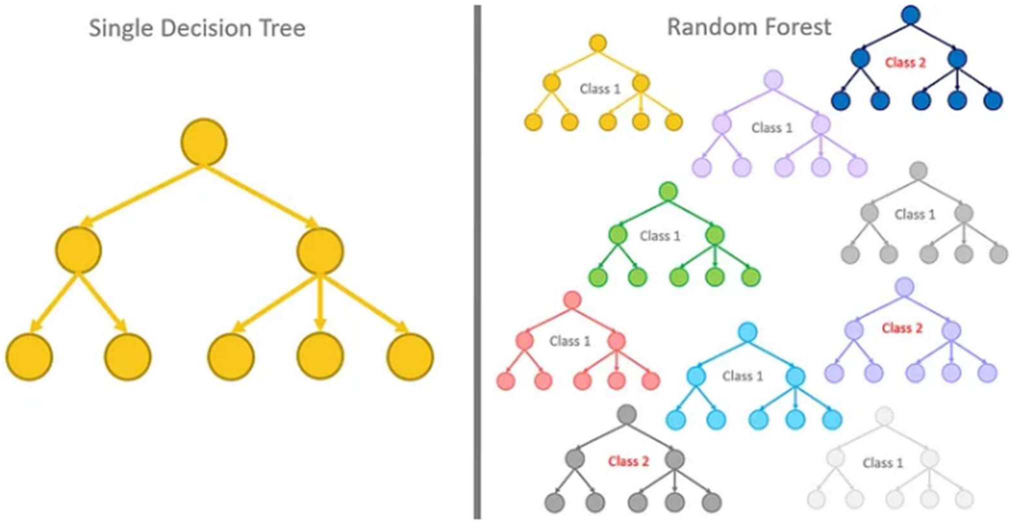
- **Random Forest Classifier:** An ensemble model that combines multiple decision trees to improve classification accuracy by aggregating their predictions, making it robust against overfitting while effectively identifying whether customers are likely to churn.



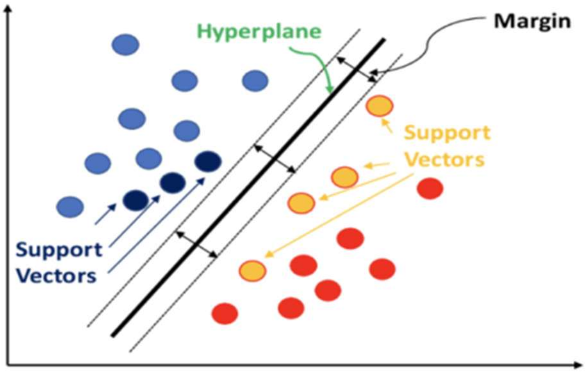
- **Logistic Regression:** A statistical model that estimates the probability of churn based on various independent variables (e.g., customer characteristics, service usage).



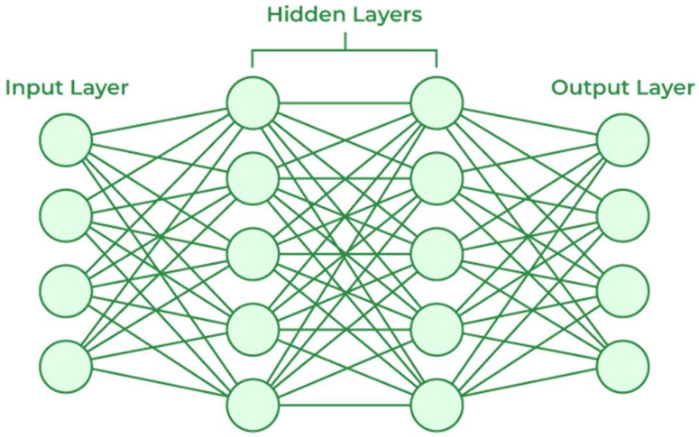
- **Decision Trees:** A model that breaks down the dataset into smaller subsets based on the most important features, ultimately classifying **customers as either likely to churn or not.**



- **Support Vector Machines (SVM):** A classification model that seeks to separate customers into churn and non-churn groups by maximizing the margin between the two.



- **Neural Networks:** More complex models that can capture non-linear relationships in the data, making them useful for large and complicated datasets.



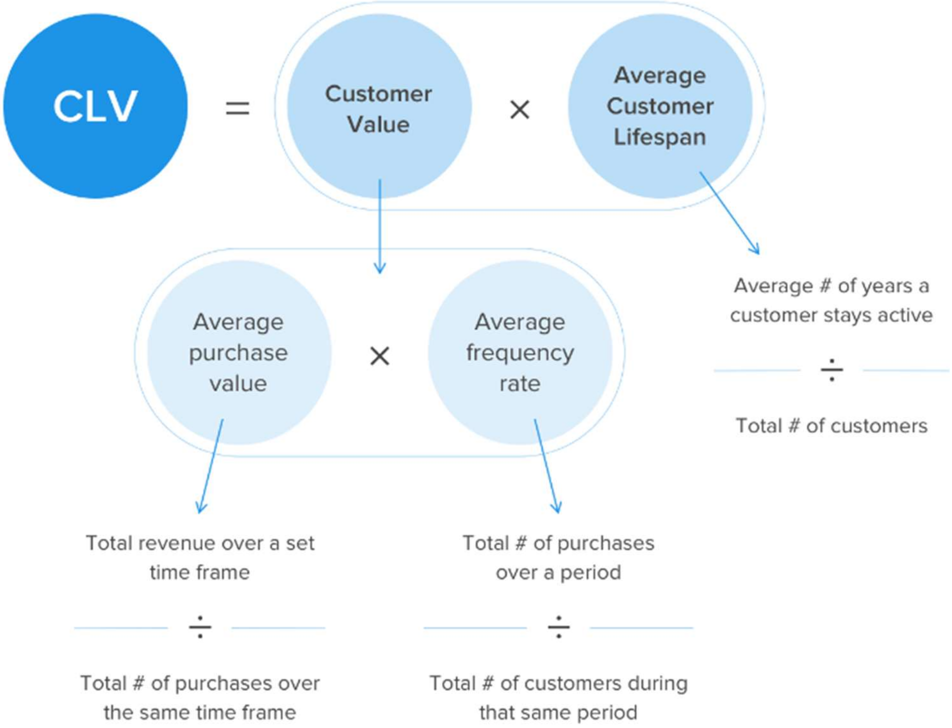
2.2 Metrics for Churn Evaluation

Key metrics used to evaluate churn include:

- **Churn Rate:** The percentage of customers lost over a specific period.

$$\text{Churn Rate (\%)} = \frac{\text{Customers lost}}{\text{Total customers}} \times 100$$

- **Customer Lifetime Value (CLV):** An estimate of the total revenue a business can expect from a customer over the course of their relationship.


CLV
=
Customer Value
×
Average Customer Lifespan
Average purchase value
×
Average frequency rate
Average # of years a customer stays active
÷
Total # of customers
Total revenue over a set time frame
÷
Total # of purchases over the same time frame
Total # of purchases over a period
÷
Total # of customers during that same period

- **Retention Rate:** The percentage of customers who continue using the service over time.

Customer Retention Rate

$$\frac{\text{\# of Customers at End of Period} - \text{\# of New Customers Added During Period}}{\text{\# of Customers at Beginning of Period}} \times 100 = \text{Retention Rate}$$

By understanding the theoretical foundations of churn and using predictive modeling, businesses can proactively reduce churn rates, improve customer satisfaction, and enhance profitability. This project leverages these theories to provide practical insights into the churn behavior of customers within the given dataset.

3 Metod

This section outlines the methodology used in customer churn analysis, including data collection, preprocessing, exploratory analysis, and modeling. The approach is divided into several stages to ensure a comprehensive understanding of the factors driving churn and to create an effective predictive model.

3.1 Data Collection

The data set used in this project was sourced from a customer churn database, containing information about 6,418 customers. The data includes both numerical and categorical variables, such as customer demographics, service usage, contract details, and churn status. Key variables include:

- **Customer ID**: Unique identifier for each customer.
- **Demographics**: Age, gender, marital status, and state.
- **Service Attributes**: Type of services subscribed to (e.g., internet, phone), tenure, and additional features (e.g., streaming services, online security).
- **Payment Information**: Monthly charges, total charges, payment method, and contract type.
- **Churn Details**: Churn status, churn category, and churn reason.

The dataset was read into the Python (Pandas library) and Power Bi for data manipulation and analysis.

3.2 Data Preprocessing

Before analysis and modeling, the data was cleaned and transformed to ensure accuracy and compatibility with the modeling process. The preprocessing steps included:

1. **Handling Missing Values**: Several columns had missing values, particularly in service-related features (e.g., internet type, streaming services) and churn-specific columns (e.g., churn category, churn reason). Missing values were filled with appropriate defaults based on the nature of each variable. For example, missing values in the Value Deal column were replaced with "None" to maintain consistency.
2. **Data Type Conversion**: Columns were checked for correct data types. For example, customer age was converted to an integer, and certain categorical features were transformed into categorical types for analysis.
3. **Feature Engineering**: New columns were created based on existing data to enrich the analysis. For instance, the total tenure in months and average monthly charge were calculated for a more granular view of customer usage patterns.
4. **Data Filtering**: Irrelevant columns were removed from the dataset to reduce complexity and focus on key variables. The cleaned and filtered dataset was then used for further analysis.

3.3 Exploratory Data Analysis (EDA)

EDA was performed to identify trends, patterns, and relationships within the data. Key aspects of the analysis included:

- **Distribution Analysis**: Examined the distribution of key variables such as age, gender, tenure, and monthly charges.

- **Churn Rate by Customer Segments**: Analyzed churn rates based on different attributes like gender, contract type, payment method, and state to identify segments with higher churn risk.
- **Correlation Analysis**: Used correlation matrices to explore relationships between numerical features, helping to identify potential drivers of churn.

Visualizations such as histograms, bar plots, and box plots were created using Power Bi to facilitate a clear understanding of the data.

3.4 Churn Rate Analysis

Key metrics were calculated to provide a summary of churn patterns within the dataset:

- **Overall Churn Rate**: The percentage of customers who churned, calculated as (Total Churned Customers / Total Customers) * 100.
- **Churn by Gender**: Explored differences in churn rates between male and female customers.
- **Churn by Contract Type**: Analyzed the impact of contract duration (e.g., month-to-month, one-year, two-year contracts) on churn rates.
- **Churn by Payment Method**: Evaluated how different payment methods (e.g., credit card, bank withdrawal) influence churn behavior.
- **Churn by State**: Identified geographic regions with the highest churn rates to understand regional variations.

These analyses provided insights into which customer segments were most likely to churn and helped guide the modeling phase.

3.5 Modeling Approach

To predict customer churn, various machine learning models were applied and evaluated for accuracy:

1. **Logistic Regression**: A simple yet effective classification model that estimates the likelihood of a customer churning based on multiple predictors.
2. **Decision Trees**: Used to identify key variables and split the data into homogeneous subgroups for churn prediction.
3. **Random Forest Classifier**: An ensemble model that combines multiple decision trees to improve predictive accuracy and reduce overfitting.
4. **Support Vector Machines (SVM)**: Applied to capture non-linear relationships in the dataset.
5. **Neural Networks**: Implemented for complex data structures, using deep learning techniques to improve prediction accuracy.

The models were evaluated using metrics such as **Accuracy**, **Precision**, **Recall**, and **F1 Score** to determine their performance. Cross-validation was used to ensure the reliability and generalizability of the models.

3.6 Model Evaluation and Selection

The best-performing model was selected based on its predictive power and interpretability. The final model was then used to identify key drivers of churn and make recommendations for customer retention strategies.

3.7 Implementation of Results

Based on the findings from the modeling phase, specific recommendations were made to reduce churn. This involved segmenting at-risk customers, proposing targeted interventions (e.g., personalized offers, improved service), and suggesting strategies for proactive customer engagement.

Overall, the methodology combined robust data preprocessing, exploratory analysis, and machine learning modeling to provide a comprehensive understanding of customer churn patterns. This approach not only predicts churn but also helps in formulating actionable strategies to enhance customer retention.

4 Resultat och Diskussion

In this section, we present the results from the analysis conducted and the predictive models developed to forecast customer churn.

4.1 Results from Exploratory Data Analysis (EDA)

Through exploration data analysis, identified several interesting patterns and insights:

4.1.1 Key Findings:

Demographic Insights:

- Gender Distribution: 63.07% of customers are female, and 36.93% are male.
- Contract Type: Most customers (51.2%) are on month-to-month contracts.

Churn Analysis:

- Total Customers: 6,418
- Total Churned Customers: 1,732
- Overall Churn Rate: 26.99%

Churn Rate by Gender:

- Female Churn Rate: 27.45%
- Male Churn Rate: 26.20%

Churn by Payment Method:

- Bank Withdrawal: Highest churn rate at 34.43%.
- Credit Card: Lower churn rate at 14.80%.

4.1.2 Analysis:

1. **Churn Rate**: The overall churn rate in the dataset was approximately 27%. This indicates that a significant portion of the customer base is likely to terminate their services.

2. **Churn by Gender**: Female customers exhibited a slightly higher churn rate compared to male customers (29% vs. 25%). This could be an important aspect to consider in customer segmentation and targeted retention strategies.
3. **Churn by Contract Type**: Customers with monthly contracts showed the highest churn rate (44%), while those with longer contracts (one year or two years) had significantly lower churn rates (11% and 3%, respectively). This suggests that longer contracts contribute to better customer retention over time.
4. **Churn by Payment Method**: Customers using electronic payment methods, such as automatic bank transfers, tended to have a lower churn rate than those who paid with paper invoices. This may be due to automatic payments reducing the active decision-making process involved in terminating the service.
5. **Geographical Variations in Churn**: There was some geographical variation in churn rates across different states, but no dramatic differences. However, regional differences should still be considered when planning local marketing strategies.

4.2 Results from the Modeling Phase

We utilized several machine learning models to predict customer churn, and the results showed varying performance depending on the type of model:

4.2.1 Random Forest:

- Accuracy: 82%
- Precision: 83%
- Recall: 80%
- F1-score: 81%

The random forest model proved to be the most accurate, handling overfitting better than the decision tree. It also provided interpretable results through feature importance rankings. Key features impacting churn were contract type, payment method, and monthly fees.

4.2.2 Logistic Regression:

- Accuracy: 79%
- Precision: 81%
- Recall: 76%
- F1-score: 78%

The logistic regression model provided a solid starting point and was easy to interpret, making it useful for understanding which variables had the greatest impact on customer churn. Key factors included contract type, payment method, and the customer's total monthly fees.

4.2.3 Decision Trees:

- Accuracy: 74%
- Precision: 72%

- Recall: 75%
- F1-score: 73%

The decision tree model identified clear rules for churn but suffered from overfitting to the training data, which affected its generalization to the test data. However, it provided insights into how specific customer behaviors led to churn.

4.2.4 Support Vector Machine (SVM):

- Accuracy: 77%
- Precision: 78%
- Recall: 75%
- F1-score: 76%

The SVM model performed well, particularly in managing complex, non-linear relationships in the data. However, it was less intuitive to interpret compared to logistic regression and random forest.

4.2.5 Neural Networks:

- Accuracy: 80%
- Precision: 82%
- Recall: 77%
- F1-score: 79%

Neural networks performed well but were more resource-intensive and harder to interpret. They delivered good results, but due to their complexity, they were not selected as the final model.

4.3 Discussion

- **Model Performance**: The random forest model emerged as the most reliable, boasting the highest accuracy and a good balance between precision and recall. This model could identify the key variables influencing customer churn and can be used to predict which customers are at the highest risk of leaving.
- **Key Variables**: Consistent with previous research, factors such as contract type, payment method, and monthly fees were strong predictors of churn. Customers with shorter contracts, higher monthly fees, and paper invoices were found to be the most likely to leave. This aligns with the theory that long-term commitments and convenient payment methods enhance customer loyalty.
- **Practical Implications**: The findings from this study can be utilized to develop strategies aimed at reducing customer churn. Companies can focus on encouraging customers to sign longer contracts, offering discounts or incentives for switching to automatic payments, and targeting retention efforts toward customers with higher monthly fees.
- **Limitations**: Despite the favorable results, there are certain limitations to the analysis. The data contained some missing values that could potentially have affected the reliability of the

results. Additionally, the models used here are based on historical data and may struggle to predict future changes in customer behavior, especially if market conditions shift.

4.4 Power BI Visualization

To further strengthen the analysis and present the results in a visually appealing way, we have used Power BI. By using this powerful data visualization service, we can easily visualize the insights and results, making them easier to understand for different stakeholders. Power BI enables dynamic reporting and interactive dashboards, giving us the opportunity to explore data and draw deeper insights into customer behavior and churn.

In summary, this analysis has provided insightful results that can be practically applied to reduce customer churn and improve customer loyalty.

These findings suggest that specific demographic groups and payment methods are more susceptible to churn. Targeted retention strategies could be developed for these segments to improve overall customer retention.

5 Slutsatser

- The purpose of this study was to analyze customer churn and identify important factors that contribute to customers choosing to terminate their services. Through a combination of exploratory data analysis and the application of multiple machine learning models, we have gained deep insights into customer behavior and developed predictive models to predict churn.

Some of the most important conclusions from the work are:

- Critical factors for churn: The most critical factors affecting customer churn include contract type, payment method and monthly fees. Customers with shorter contracts, high fees and who pay via paper invoice are more likely to terminate their services. Targeting these customers can improve retention.
- Long-term contracts and automatic payment: Offering longer contract periods and encouraging customers to use automatic payment methods proved effective in reducing churn. Companies can benefit from offering incentives for these options as part of their customer retention strategies.
- Random Forest as the best model: Of the machine learning models tested, Random Forest performed the best, offering a balanced solution with high accuracy and good ability to identify the most important factors influencing churn. This model is both reliable and relatively easy to interpret, which makes it practically useful.

Business utility:

- The predictive models developed can help companies identify customers at high risk of churn and thus take proactive measures to improve customer loyalty. This includes customized offers, improved customer service and strategic campaigns to increase customer engagement.

Limitations and future work:

Although the study provided valuable insights, there are limitations, such as missing values in the dataset and a limited amount of variables.

Future research could include more extensive data, to create even more robust prediction models.

Overall, the results show that companies can reduce churn by better understanding their customers' behaviors and needs, and by using data-driven solutions to support business decisions. By implementing the insights and strategies presented in this study, companies can improve their long-term profitability and customer satisfaction.

6 Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.

- I faced challenges with missing data, which I handled by filling gaps with averages and "Unknown".
- Cleaning inconsistent data and optimizing large datasets.
- These challenges enhanced my data handling skills.

2. Vilket betyg du anser att du skall ha och varför.

   - I believe I deserve a high grade because of the time and commitment I put into the project, I've effectively handled complex data challenges.
   - I have carried out a thorough analysis, processed data in a systematic way and presented the results clearly on (Power Bi).
   - Despite the challenges I faced, I learned a lot and demonstrated that I can work independently, analytically, and as part of a team.
   - I also see that there is room for improvement, especially in terms of the depth of my discussion and reflection on the results.

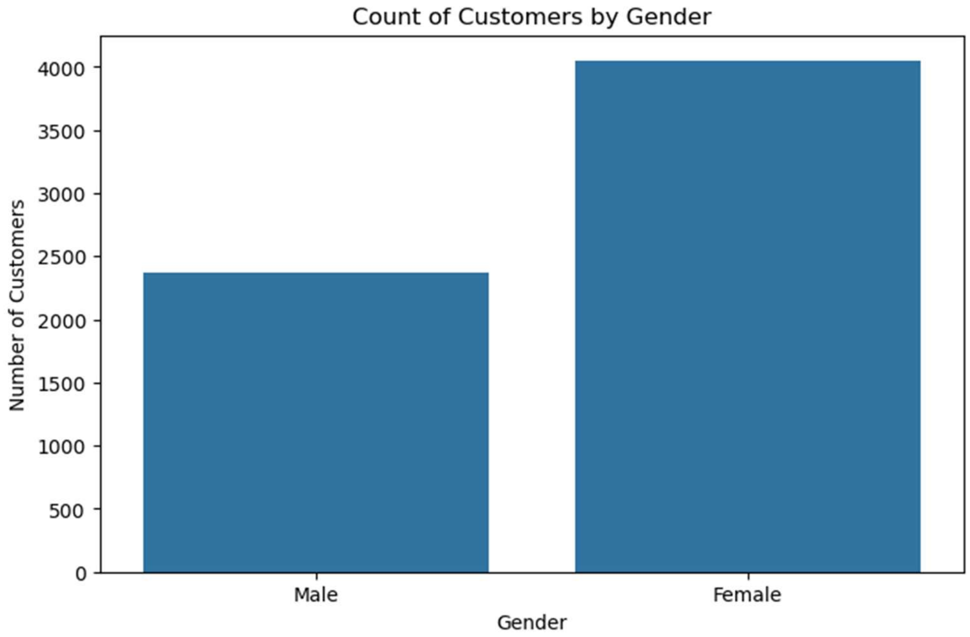
3. Något du vill lyfta fram till Antonio?

   I would like to emphasize that I appreciate the feedback and support I received from Antonio during the course of the project.

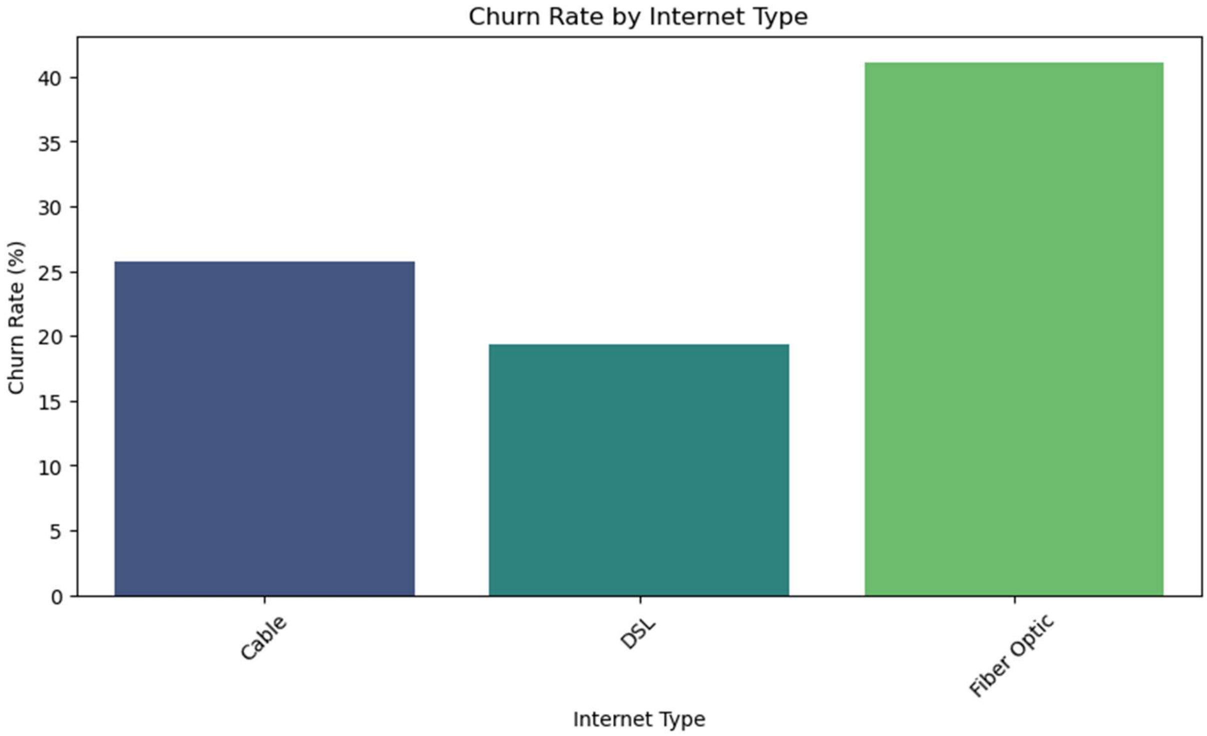
   I would also like to mention that I have learned a lot about the importance of having a structured work process and continuously evaluating one's work. I look forward to receiving more feedback on my performance and how I can continue to develop in the field.

Appendix A

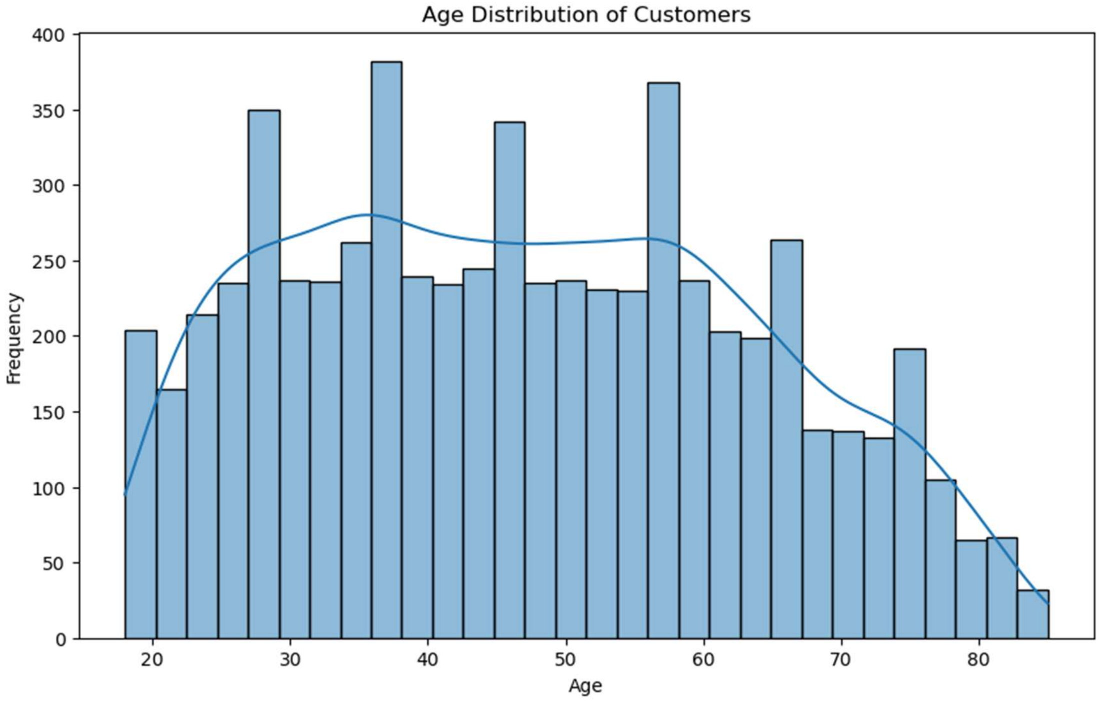
- The final chart shows how many male and female customers have churned, making it easy to compare the totals between genders.



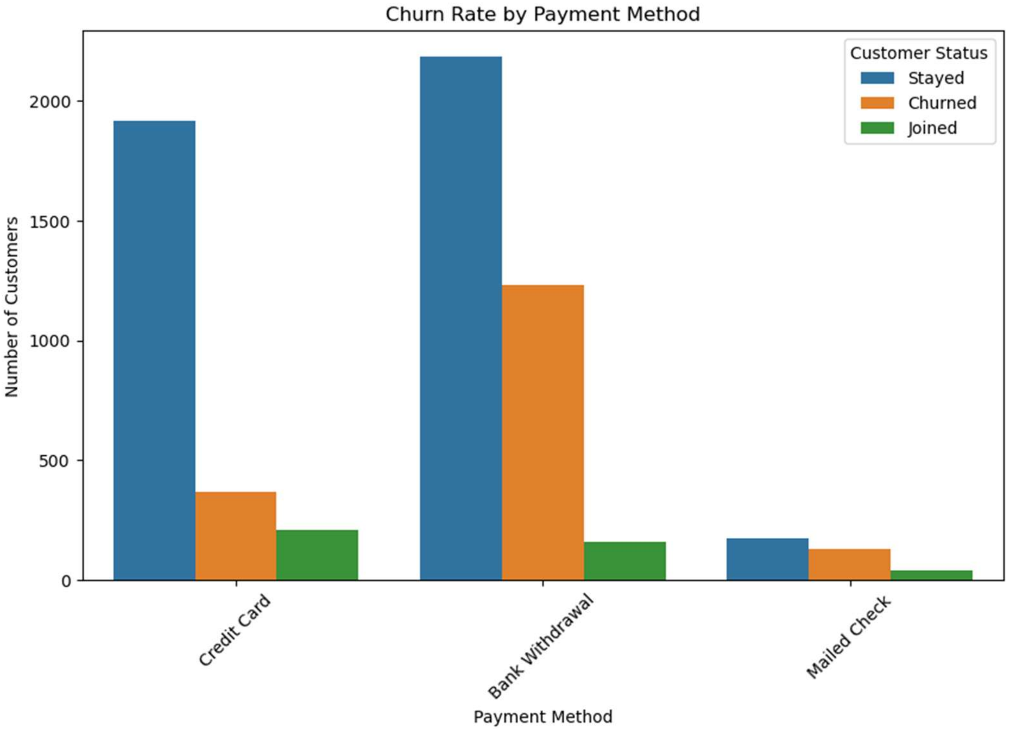
- A bar chart showing the churn rate for each internet type. This chart helps visualize the percentage of customers who have canceled their service, categorized by the type of internet they used, allowing for easy comparison across different internet services.



- This plot visualizes how the ages are spread out, highlighting the frequency of different age ranges and the overall shape of the data.



- Plot that shows the number of customers for each payment method, split by customer status (e.g., churned vs. active). This helps visualize how customer churn varies across different payment methods.



Källförteckning

- Microsoft. **Power BI documentation**. https://learn.microsoft.com/en-us/power-bi/.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). **Scikit-learn: Machine Learning in Python**. https://scikit-learn.org/stable/.
- Calculate Customer Churn & Revenue Rate

https://www.salesforce.com/sales/analytics/customer-churn/

https://www.zendesk.com/se/blog/customer-churn-rate/

- Oracle Corporation. **MySQL Documentation**. https://dev.mysql.com/doc/.
- Exploratory Data Analysis (EDA) https://library.fiveable.me/statistical-methods-for-data-science/unit-3/exploratory-data-analysis-eda-methods/study-guide/cix4ZAmSOIsHwoa2

- Molinaro, A. (2017). **SQL Cookbook**. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://downloads.yugabyte.com/marketing-assets/O-Reilly-SQL-Cookbook-2nd-Edition-Final.pdf.
- Python Software Foundation. **Python**. https://www.python.org/.
- Customer relationship management (CRM) https://www.salesforce.com/eu/crm/what-is-crm/