Project report

Why do customers churn?

Author: Thea Törnqvist
Date: 2025-01-10

Table of contents

1. Identifying datasets

For this project, I have chosen to analyze customer churn using data from a US-based telecom company. The analysis is based on six interconnected datasets from the same company, each providing unique insights into customer behavior. These datasets include:

- **Customer Info**: Demographic data.
- **Location Data**: Geographic information.
- **Service Options & Online Services**: Details of subscribed services.
- **Payment Info**: Billing and payment information.
- **Customer Churn**: Churn status and reasons.

These datasets are particularly valuable for understanding customer behavior and supporting business decision-making. Customer churn is a critical issue for companies, making this a realistic and impactful case study. By analyzing how demographics, service choices, and payment behaviors influence churn, the project demonstrates how data-driven decisions can improve customer retention and satisfaction. Personally, as someone aspiring to work with both people and data, this analysis aligns with my professional interests and showcases how businesses can leverage data to make informed decisions.

To extract meaningful insights, I merged all six datasets using the common identifier *customer_id*. Before merging, I conducted several checks and cleaning steps:

- **Column Names**: Ensured consistency across datasets.
- **Missing Values**: Identified missing values in the churn reason column and left them unchanged, as they might still provide valuable insights.
- **Duplicate Columns**: Found duplicates in the columns *phone_service* and *internet_service*. Upon confirming they were identical, I retained only one of each.
- **Verification**: Checked the merged dataset using *merged_data.info()* to ensure no data was lost or misaligned.

By cross-referencing these datasets, I could analyze how combinations of factors – such as service usage, payment methods, and demographic attributes – impact customer churn. This integrated approach provided deeper and more actionable insights than analyzing the datasets individually.

2. Visualization and descriptive statistics

To better understand customer churn and underlying patterns, I divided the visualizations into sections for a clearer overview of key variables.

2.1 Descriptive statistics

Using the *.describe()* function, I analyzed the mean, standard deviation, range, and quantiles of selected numerical variables. Given the dataset's size, I focused on variables most relevant to understanding churn.

- **Age:**
  Age influences customer behavior, preferences, and loyalty. Younger customers may switch services more easily, while older customers prioritize stability. The average age is 46 years, with a wide distribution (19–80 years), slightly skewed toward younger customers.
- **Tenure:**
  Tenure reflects customer loyalty. The average tenure is 32 months (~2.7 years), with 25% of customers staying for 9 months or less—indicating a higher churn risk among newer customers.
- **Total Population (Per Zip Code):**
  Population density may affect service availability and competition. The average population per zip code is ~22,139, but the large standard deviation (~21,152) highlights significant variation, ranging from 11 to 105,285. Urban areas may face higher churn due to more competition, while rural areas may have more stable customers.
- **Satisfaction Score:**
  This score directly reflects perceived service quality. The average score is 3.24, suggesting neutral satisfaction. Customers scoring 1–2 (under 25%) are at higher churn risk, while most customers are neutral or satisfied.
- **Total Charges:**
  Payment amounts can impact retention. The average total charge is ~$2,280, with a wide range ($18.80–$8,684.80), reflecting diverse spending habits. Lower charges may indicate new customers, while higher charges could suggest long-term customers, potentially at risk of churn if they perceive poor value.

By focusing on these key variables, we gain a clearer understanding of the factors driving customer churn and can better target strategies to improve retention.

2.2 Demographic and geographic variables

In this section, I chose to do a 2x3 grid with six different demographic and geographic variables that I thought could be meaningful in understanding the customers and overall data.

The **age** distribution skews slightly younger, with most customers between 20 and 60 years old, which emphasizes the analysis from the descriptive statistics. Younger customers may be more likely to switch providers for better deals, while older customers might value service stability. When it comes to **gender**, the near-equal split between **male and female** customers suggests broad service appeal. However, exploring gender-specific churn patterns could reveal opportunities for more personalized marketing and retention efforts. **Married** and **unmarried** customers are almost evenly split. Married customers may favor bundled family plans, promoting loyalty, while unmarried customers might prefer flexible, individual services, possibly increasing churn risk.

Most customers do not have **dependents**, indicating a younger or single demographic that may prioritize flexible and affordable plans. Customers with dependents might value stability and comprehensive service bundles, potentially lowering their churn risk. When it comes to the **number of dependents**, the distribution is heavily skewed toward zero, indicating that most customers have no dependents, which aligns with a younger customer base. Customers with dependents might have different service needs (e.g., family plans, security features) and could be more loyal due to the inconvenience of switching providers.

Lastly, the **total population** distribution is highly skewed, spanning from rural to densely populated urban areas which also emphasizes the analysis of the descriptive statistics.

2.3 Service subscription

In this section, I analyzed the distribution of various service subscriptions to understand customer preferences and potential churn risks.

The **phone service** and **internet service** have the highest subscription rates, with the majority of customers using these essential services. This reflects their foundational role in telecom service offerings. However, the slight drop in internet service subscriptions compared to phone service may indicate opportunities to improve internet packages.

In contrast, optional services like **online security**, **online backup**, **device protection**, and **premium tech support** show significantly lower subscription rates. These services have a near 30/70 split between subscribers and non-subscribers. This suggests that customers may not perceive enough value in these add-ons or may be deterred by additional costs. Improving the perceived value or bundling these services with core offerings could help increase adoption and reduce churn.

The **multiple lines** service also shows moderate adoption, indicating that a sizable portion of customers either do not need additional lines or prefer more flexible, single-line plans. Customers with multiple lines may have higher switching costs, making them more loyal, while single-line users could be more prone to churn.

Understanding these service adoption patterns helps identify areas for targeted marketing, product improvement, and customer retention strategies.

2.4 Payment and contract behavior

The analysis of payment methods, contract types, total charges and tenure highlights key patterns in customer behavior and potential churn risks.

For different **payment methods** we see that electronic checks (33.6%) are the most common but may indicate higher churn risk due to their manual nature, leading to missed or late payments. Mailed checks (22.9%) also reflect a preference for traditional, manual billing, which could correlate with increased churn. Automated payments via bank transfer (21.9%) and credit card (21.6%) are less common but typically suggest more stable, loyal customers due to the convenience of automation.

When it comes to **contract type,** month-to-month contracts (51.3%) dominate, offering flexibility but posing a higher churn risk due to the ease of switching providers. One-year (22.0%) and two-year (26.7%) contracts imply greater customer commitment and reduced churn due to contract terms and incentives.

The **Total Charges** distribution is heavily right-skewed, with most customers paying under $2,000, indicating many are new or low-usage users. In contrast, fewer customers have high charges, likely reflecting long-term, loyal users. The **Tenure** distribution also skews right, with many customers leaving within the first 10 months, while a peak at 72 months suggests strong loyalty among long-term subscribers. Both of these analyses emphasize the analysis of the descriptive statistics.

2.5 Churn analysis

In this final section, I focused on visualizing the churn rate to highlight the proportion of customers who have left the company versus those who remained. The visualization shows that while the majority of customers stayed, around **1,800 customers** churned—accounting for approximately **25%** of the customer base. This considerable churn rate indicates that customer retention is a significant challenge for the company and warrants further investigation to identify and address the underlying causes.

3. Formulating hypotheses

To better understand the drivers behind customer churn, I have formulated three hypotheses based on observed trends and logical reasoning. Each hypothesis targets a specific factor that could influence churn behavior, providing valuable insights for improving customer retention strategies.

Hypothesis 1: Customers with month-to-month contracts are more likely to churn.

Month-to-month contracts offer customers greater flexibility to cancel services without financial penalties, making them more susceptible to switching providers for better deals or improved services. In contrast, long-term contracts (one-year or two-year) typically involve commitments or early termination fees, discouraging customers from leaving. Investigating this relationship is important because it can help the company determine whether promoting longer contracts or offering incentives for contract renewals could effectively reduce churn. If this hypothesis is confirmed, the company could implement strategies to convert month-to-month customers into long-term subscribers through discounts, loyalty programs, or exclusive service bundles. This approach could enhance customer retention and stabilize revenue streams.

Hypothesis 2: Customers in urban areas are more likely to churn.

Urban customers typically have access to a wider range of service providers due to higher market competition in densely populated areas. This increased competition gives urban customers more options to switch to competitors offering better pricing, faster internet, or more attractive service bundles. In contrast, rural or suburban customers might face limited provider options, making them less likely to churn.Validating this hypothesis would help the company focus its customer retention efforts in competitive urban markets. Strategies such as personalized marketing, improved service quality, or targeted promotions could be deployed in these areas to reduce churn and strengthen market share.

Hypothesis 3: Customers with more services are less likely to churn.

Customers who subscribe to multiple services (e.g., phone, internet, streaming, device protection) are likely more invested in the company's ecosystem. Bundling multiple services typically increases switching costs due to the inconvenience of canceling and setting up services with another provider. Additionally, customers with more services may perceive greater value, increasing their satisfaction and loyalty. If this hypothesis holds true, the company could focus on upselling and bundling services to increase customer engagement and reduce churn. Promotions that encourage customers to add services or create customized bundles could improve satisfaction and make it harder for competitors to lure customers away.

4. Statistical analysis

4.1 Hypothesis 1

Hypothesis: Customers with month-to-month contracts are more likely to churn.

H_0: $Churn\ rate\ is\ independent\ of\ contract\ type.$

H_0: $P_{Month-to-Month} = P_{One-Year} = P_{Two-Year}$

H_a: $Churn\ rate\ depends\ on\ contract\ type.$

H_a: $P_{Month-to-Month} \neq P_{One-Year} \neq P_{Two-Year}$

4.1.1 Method

To investigate whether customers with **month-to-month contracts** are more likely to churn, a **Chi-square test** was used. This statistical test is ideal for examining relationships between two categorical variables—in this case, the **contract type** (Month-to-Month, One-Year, Two-Year) and the **churn label** (Yes, No). The Chi-square test works by comparing the observed frequencies of customers in each contract category who churned or stayed with the expected frequencies if no relationship existed.

The test's assumptions were carefully checked. Each customer is associated with only one contract type and has a single churn status, ensuring the independence of observations. Both variables are categorical, and the expected frequencies in all categories were above 5, meeting the requirements for the test.

A statistically significant result would indicate that churn behavior differs across contract types. Specifically, if **month-to-month customers** are more prone to churn, it may suggest that the flexibility of these contracts makes it easier for customers to leave. This insight could help the company adjust its contract strategies to improve customer retention.

Contract type	Churn = Yes	Churn = No	Total
Month-to-Month	1655	1955	3610
One-Year	166	1384	1550
Two-Year	48	1835	1883
Total	1869	5174	7043

Table 1. Contingency table for churn_label and contract

Above is the contingency table to show the distribution of churn across contract types. We can intuitively see that we should get expected frequencies larger than 5. However, to be completely sure we verify it with the *print(expected)* in Python. Then, we perform a chi-square test to evaluate the independence of churn and contract type.

4.1.2 Results

We can see that the chi-square statistic is **1445.29**, which indicates that the observed churn rates for different contract types deviate from the expected churn rates. The p-value is 0. Since the **p-value is less than 0.05, we reject the null hypothesis**. In conclusion, there is a statistically significant relationship between contract type and churn. From the bar chart, we can see that month-to-month contracts have a much higher churn rate compared to one-year and two-year contracts. Two-year contracts have the absolute lowest churn rate. These results support the business intuition that customers with shorter commitments are more likely to churn, whereas long-term contracts promote customer retention.

4.2 Hypothesis 2

Hypothesis: Customers in urban areas are more likely to churn.

H_0: $Churn\ rate\ is\ independent\ of\ whether\ the\ customer\ resides\ in\ an\ urban\ or\ rural\ area.$

H_0: $P_{Urban} = P_{Rural} = P_{Suburban}$

H_a: $Churn\ rate\ depends\ on\ whether\ the\ customer\ resides\ in\ an\ urban\ or\ rural\ area.$

H_a: $P_{Urban} \neq P_{Rural} \neq P_{Suburban}$

4.2.1 Method

The original total_population variable is continuous, making it difficult to interpret how population size directly impacts churn. By grouping the population into clear categories – **Rural** (<10,000), **Suburban** (10,000–50,000), and **Urban** (>50,000) – I simplified the data and made it easier to compare churn rates across distinct area types. This approach allows for clearer insights into whether customers in more densely populated areas are more prone to churn.

To explore whether customers in urban areas are more likely to churn than those in suburban or rural areas, a **Chi-square test** was also applied. As mentioned before, this test effectively identifies relationships between two categorical variables—here, **area type** (Urban, Suburban and Rural) and **churn label** (Yes, No). The purpose was to determine whether customer churn rates vary across different geographic regions.

Area type	Churn = Yes	Churn = No	Total
Rural	694	2152	2846
Suburban	955	2514	3469
Urban	508	220	728
Total	5174	1869	7043

Table 2. Contingency table for churn_label and area_type

Similar to previously, we can see from the contingency table above that the expected frequencies should be larger than 5, which is also validated in Python code. Just looking at the table, it seems like the proportions of churned are quite similar for all area types. We can also see that rural and suburban areas have significantly more customers compared to urban areas, which might influence the results.

If the test results are statistically significant, it would suggest that the likelihood of churn differs by area type. For example, a higher churn rate in **urban areas** could be due to greater competition or easier access to alternative service providers. Understanding this dynamic can help the company tailor its marketing and service strategies to different geographic markets.

4.2.2 Results

Here the chi-square statistic is 13.58 and the p-value is 0.0011. Since $0.0011 < 0.05$ **we can reject the null hypothesis**, which means that there seems to be a significant relationship between area type and churn rates. Even though the p-value shows significance, the chi-square value suggests that the difference is not very large.

In our bar chart, we can observe that urban areas have a somewhat slightly higher churn rate compared to suburban and rural areas. The higher churn rate in urban areas could be influenced by competition or other factors specific to these regions. However, the difference is subtle and the result suggests that area type is not a very strong predictor of churn on its own.

4.3 Hypothesis 3

Hypothesis 3: Customers with more services are less likely to churn.

H_0: $Churn\ rate\ is\ independent\ of\ the\ number\ of\ subscribed\ services.$

$H_{0:}\ \mu_{Churned} = \mu_{Not\ Churned}$

H_a: $Churn\ rate\ depends\ on\ the\ number\ of\ subscribed\ services.$

$H_{a:}\ \mu_{Churned} \neq \mu_{Not\ Churned}$

4.3.1 Method

To assess whether customers with more subscribed services are less likely to churn, I created a num_services variable to count the number of services per customer. For the testing, a **Welch's t-test** was conducted. This test is designed to compare the mean number of services between customers who churned and those who did not. Unlike the standard t-test, Welch's t-test is more robust because it does not assume equal variances between groups, making it well-suited for this analysis since the two groups may have different sample sizes and variability.

Key assumptions for the Welch's t-test were satisfied. The **number of services** is a continuous variable, and each customer's churn status is independent. Since the test allows for unequal variances, it was appropriate for comparing the average number of services across the two groups.

A significant result, particularly if churned customers have a lower average number of services, would suggest that offering **more bundled or additional services** could increase customer retention. This insight emphasizes the importance of engaging customers through comprehensive service offerings.

4.3.2 Results

The Welch t-test yielded a t-statistic of -5.95 and a p-value of 2.80×10^{-9}. Since the **p-value < 0.05, we reject the null hypothesis** - indicating a statistically significant difference in the number of services between churned and non-churned customers. In our case, the negative t-value suggests that the churned group has fewer services on average compared to the non-churned group. Customers who churn have a lower mean number of services compared to those who stay. The boxplot supports the statistical test, where we can observe that the median for churned customers (orange) is around 3 services but with a

narrower distribution compared to non-churned customers (green), that also has a median around 3 services but with a broader distribution.

4.4 Summary

Hypothesis	**Statistic (T- or chivalue)**	**P-value**	**Reject H_0**
1: Contract type	1445.29	0	Yes
2: Area type	13.58	0.0011	Yes
3: Number of services	-5.95	2.80×10^{-9}	Yes

Table 3. Summarizing the statistical test results for each hypothesis

The table above provides a summary of the statistical test results for each of the three hypotheses. In conclusion, we can see that all three of them are statistically significant (at a 5% significance level). For **Hypothesis 1** (Contract type), the results strongly indicate a relationship between contract type and churn, suggesting that month-to-month contracts are associated with higher churn rates. While **Hypothesis 2** (Area Type) displays that a statistically significant relationship exists between area type and churn, the chi-square statistic and bar plot suggest that the difference is relatively subtle. For **Hypothesis 3** (Number of services), there is strong evidence that the number of services is significantly lower for churned customers compared to non-churned customers, supporting the hypothesis that customers with more services are less likely to churn.

5. Machine learning predictions

The objective of this section is to predict customer churn using machine learning models. This involves determining whether a customer will remain with or leave the company based on various features in the dataset. Understanding the key factors influencing churn will help companies implement targeted strategies to improve customer retention. Additionally, comparing model performance will reveal which model provides the highest accuracy and serves as the most reliable predictor for churn. Predicting customer churn is crucial because retaining customers is more cost-effective than acquiring new ones. By accurately identifying customers at risk of churning, the company can implement targeted strategies, reduce revenue loss and improve customer satisfaction. This task directly supports the business goal of minimizing churn and maximizing customer lifetime value.

The following research questions will guide this analysis:

RQ1: *Which features are the most significant predictors of customer churn?*

RQ2: *How accurately can machine learning models predict whether a customer will churn?*

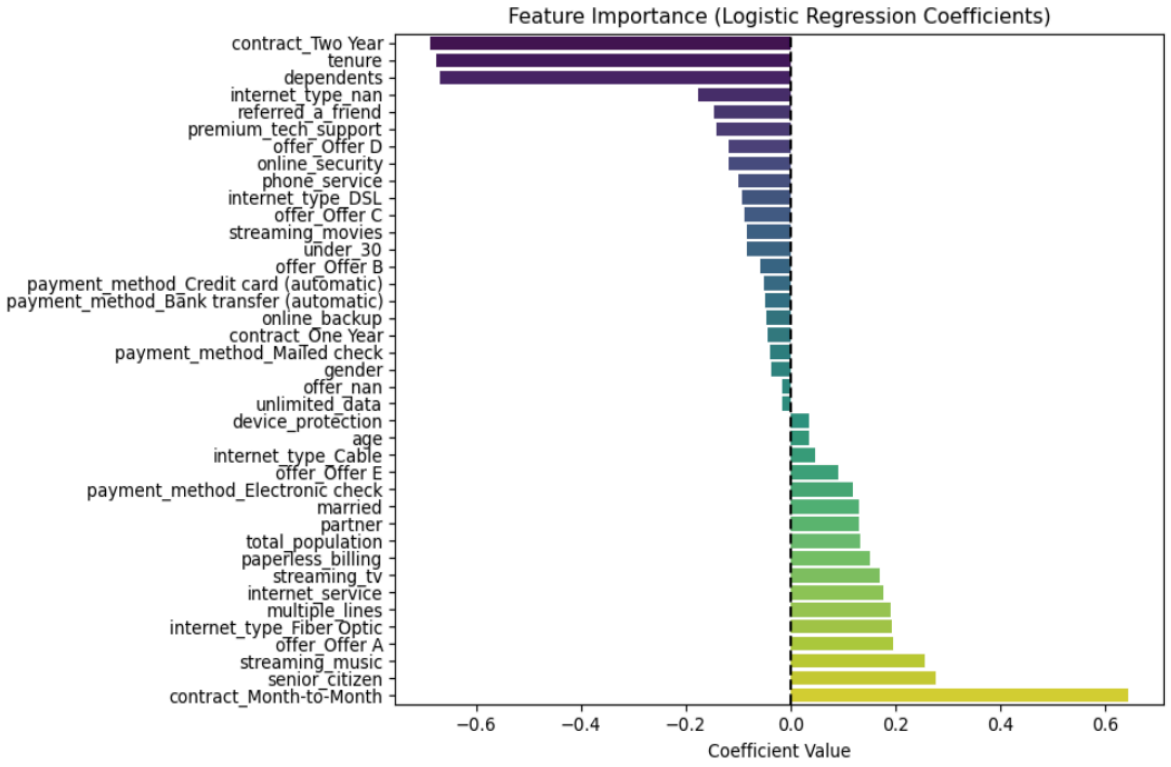
RQ3: *Which machine learning model performs best for predicting churn, and why?*

For this analysis, I have selected **Naive Bayes**, **Logistic Regression**, and **Decision Tree** models to predict customer churn. These models were chosen due to their complementary strengths in handling classification tasks, providing both interpretability and predictive performance. By combining these three models, I can compare their predictive performance and gain insights from their different approaches. Naive Bayes provides a probabilistic baseline, Logistic Regression offers interpretability with linear decision boundaries, and Decision Trees capture non-linear relationships and feature interactions.

5.1 Choosing features

For effective churn prediction, feature selection is essential to avoid overfitting, complexity, and reduced model performance. Using all features would have introduced noise, increased computational demands, and made the model harder to interpret. Overfitting, in particular, would cause the model to perform well on training data but poorly on new data.

However, I had a hard time deciding which features I wanted to use. To address this, I first analyzed feature importance using around 25 features, to identify the most impactful variables. I excluded all features from the churn_reason dataset to prevent data leakage, as these variables directly relate to the target (churn_value) and would falsely inflate model accuracy.



Graph 1. Feature Importance (Logistic Regression Coefficients)

As shown in the graph below, we can see that some of the most important features include *tenure*, *contract type*, *dependents*, *internet type* and *senior citizens*. I then combined these with the remaining features from the hypothesis testing that I did not include in the initial testing – *number of services* and *area type* – which were statistically proven to influence churn and to build on the hypothesis testing from earlier. By refining the feature set to include only relevant and significant variables, the models are more accurate, efficient, and interpretable.

5.2 Train-test split and preprocessing

Train-test split

After importing necessary modules, I defined the selected features (X) and the target variable (y). Then for the target variable, I chose "churn_value" which again is a binary variable with either 1 (Yes) or 0 (No). I also chose to use a 80/20 train-test split since that is a common split that ensures enough training data. Random state 42 is also commonly used and has no meaningful purpose in our case. The stratify parameter ensures that the proportion of classes in the target variable is preserved in both the training and testing datasets so that it doesn't bias the model with different ratios. Lastly, I wanted to double-check the shape of the datasets to make sure it looked good, which it did.

Encoding categorical variables

Since machine learning models can't handle text data directly, we need to handle our categorical data by encoding them. For binary categories like *dependents* and *senior citizen*, I use ***Label Encoding*** to give them binary values. When it comes to categories with more than two values – such as *contract*, *area type* and *internet type* – I have used ***One-Hot Encoding*** so every category becomes separate columns - which prevents the model from assuming any order in the categories.

Important to think about at this stage is that we want to **fit the training dataset only and transform both** to avoid data leakage. If we fit it on the entire dataset (both train and test) the encoder "sees" the test dataset which can lead to overfitting and will give a false sense of model performance.

First, I identified which columns were binary and which were multi-category columns. I started off by encoding the binary ones with LabelEncoder() and then moved on to the multiple category variables with OneHotEncoder(). Lastly I created data frames that contained the correct column names, because I noticed that if I didn't, the columns turned into indexes instead. I used concat to merge the coded columns with the remaining data and printed it to make sure it looks correct. Lastly, to double-check the data, I printed the shape, dtypes, columns, datatype and sum of columns to make sure that the data looks good.

Feature Scaling

After encoding, we also want to scale our data to make sure that all features contribute equally to the model. If we don't, the model might prioritize the larger-scaled feature. Since we are using models like logistic regression and Naive Bayes, feature scaling is important for proper model performance. Models like decision trees don't need scaling, but it doesn't "hurt" to do it anyways. Logistic regression however, assumes features are on similar scales for proper convergence. Naive bayes also work better if features are normally distributed and scaled.

In this case I used StandardScaler() which standardized features by removing the mean and scaling to unit variance, which means that all features will have a mean of 0 and a standard deviation of 1 after scaling.

5.3 Training and evaluating models

5.3.1 Evaluation metrics

I have chosen to use Naive Bayes, Logistic Regression and Decision Tree to make predictions for customer churn. For every model, we receive accuracy score, recall, precision and F1-score to analyse the results. I also visualize the confusion matrix for extra clarity for each model.

Accuracy measures the proportion of correct predictions out of all predictions made. If the accuracy score is high (close to 1) it means that most predictions (for both churners and non-churners) were correct. However, in imbalanced datasets, like the dataset I'm using, accuracy can be misleading since the model can predict the majority class well while failing to predict the minority class. For example, if most customers don't churn, then predicting "no churn" for everyone could give high accuracy but fails to identify churners.

Precision measures how many of the predicted positive cases were actually positive. It calculates the true positives divided by all positives. If we receive a high precision value (close to 1) it means that when the model predicts churn, it's usually correct. Here the focus lies on reducing false positives (predicting churn when it's actually not churn).

Recall measures how many of the actual positive cases were correctly identified. In contrast to precision, the focus is on reducing false negatives (missing customers who will churn). If we have a high recall value, it ensures that the model catches most customers who are likely to churn.

The **F1-score is** the harmonic mean of precision and recall. It balances the trade-off between them. A high F1-score means that the model balances precision and recall well.

The **confusion matrix** shows the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN).

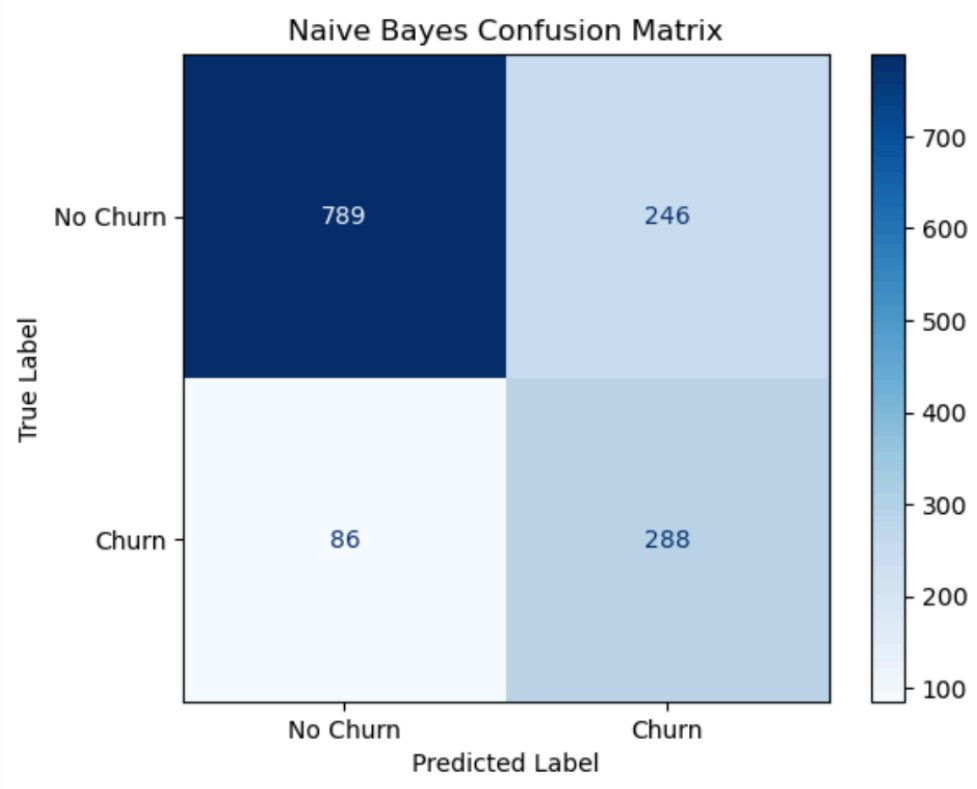
5.3.2 Naive Bayes (NB)

Naive Bayes is a probabilistic classifier based on Bayes Theorem. It predicts the probability of a data point belonging to a particular class. It is a simple, fast model that handles categorical data well that assumes feature independence. To implement this model I used GaussianNB from the sklearn package and first trained it on the training data and used the testing data to make predictions.

	Precision	Recall	F1-score
0 (Non-churn)	0.90	0.76	0.83
1 (Churn)	0.54	0.77	0.63
Accuracy			0.76

Table 4. Naive Bayes performance on customer churn

The Naive Bayes model is highly accurate when predicting non-churn; there is a high precision (90%), it correctly identifies 76% of actual non-churners and the balance between precision and recall is strong (0.83). However, for predicting churn, only 54% of predicted churn cases were correct. It only catches 77% of actual churners which is a decent recall but the balance (F1-score) was weaker than for non-churn. Overall, the precision for class 1 (churn) is fairly low.



Graph 2. Naive Bayes Confusion Matrix

The confusion matrix above emphasises that the model is good at detecting churners (77%) and correctly identifies most customers who stay. However, it shows many false positives, meaning too many non-churners are flagged as churners. Also, it had 85 missed churners (FN), and failing to catch these can be costly.

5.3.4 Logistic Regression

Logistic regression is a supervised machine learning algorithm used for classification tasks. It predicts categorical outcomes (e.g. churn or not churn) and also estimates probability that a given input belongs to a particular class. It models the relationship between the input features and the probability of a binary outcome using the logistic function. Unlike Naive Bayes, it is not as sensitive to feature correlation. To implement the model I used LogisticRegression in sklearn and used a maximum iteration of 1000.

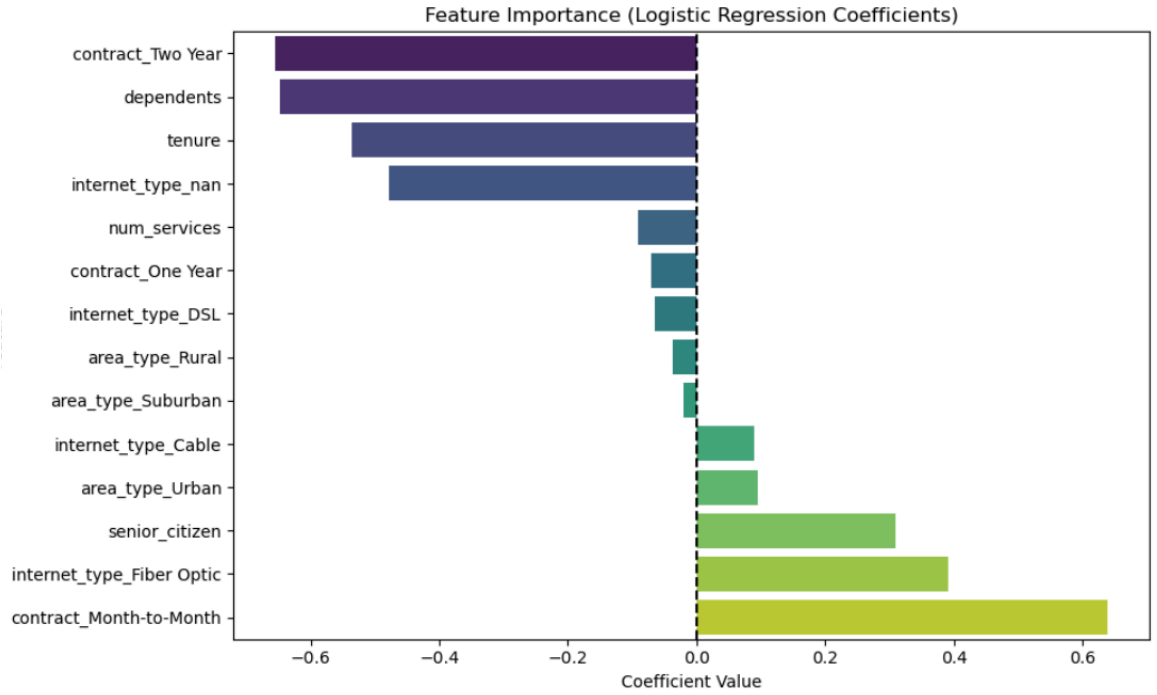
	Precision	**Recall**	**F1-score**
0 (Non-churn)	0.87	0.88	0.87
1 (Churn)	0.65	0.62	0.64
Accuracy			0.81

Table 5. Logistic regression performance on customer churn

Looking at the results from the logistic regression model, it shows stronger and more balanced performance compared to the Naive Bayes model. For non-churn customers, the model achieves a high precision of 87% and recall of 88%, resulting in a solid F1-score of 0.87. This means that the model is highly effective at correctly identifying customers who will stay with the company.

For churners, the logistic regression model performs better than Naive Bayes, with a precision of 65% and a recall of 62%. Although the recall is slightly lower than Naive Bayes, the precision is notably higher. This balance results in a better F1-score (0.64) for churn predictions compared to Naive Bayes. Overall, the model achieves an accuracy of 81%, outperforming Naive Bayes in overall prediction correctness.

I also chose to plot the feature importance from the logistic regression to evaluate what features impacts the churn value the most.



Graph 3. Feature Importance (Logistic Regression Coefficients)

From the plot we can observe that **Month-to-Month contracts**, **Fiber Optic internet**, and being a **Senior Citizen** significantly increase churn risk, indicating that flexible contracts and certain demographics are more prone to leaving. In contrast, **Two-Year contracts**, having **Dependents**, and longer **Tenure** reduce churn likelihood, reflecting stronger customer loyalty. These findings align with the hypotheses, emphasizing that contract type, service engagement, and customer demographics are important for predicting churn and guiding targeted retention strategies.

5.3.5 Decision Tree

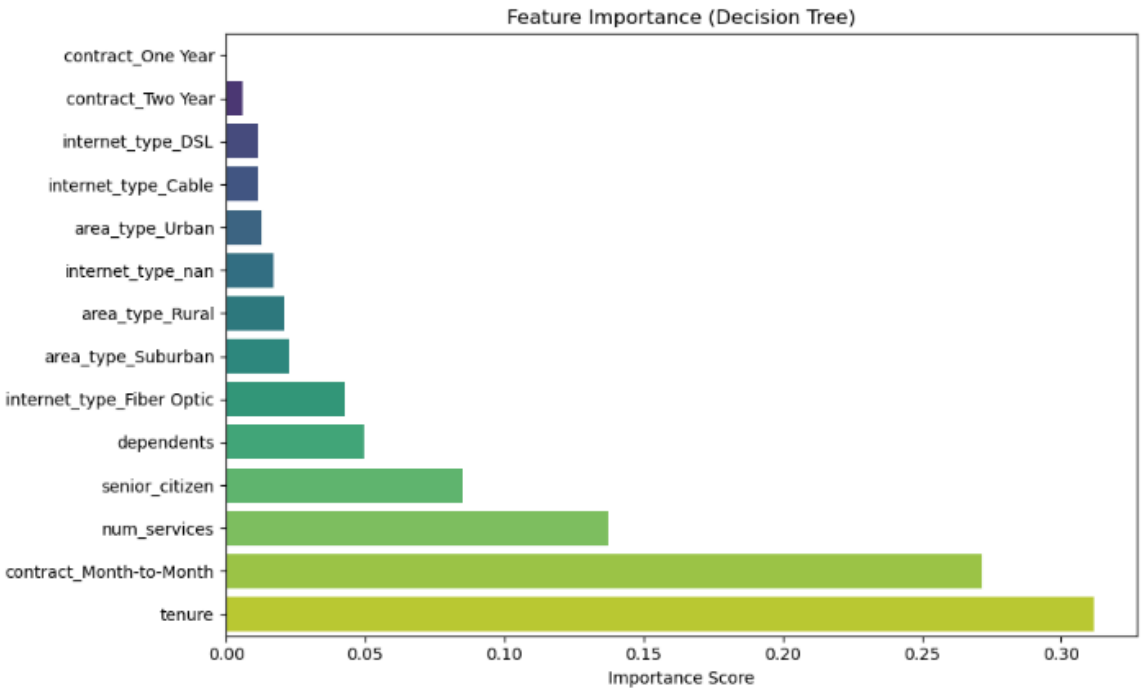
A decision tree is a supervised machine learning algorithm used for both classification and regression tasks. It works like a flowchart, where each internal node represents a decision on a feature, each branch represents the outcome of the decision and each leaf node represents the final prediction. The tree splits the data into subsets based on the feature that best separates the target classes. The algorithm evaluates every feature and threshold to decide how to split the data. It selects the split that maximizes class separation. The process repeats recursively, creating more splits until a stopping condition is met. A new sample flows through the tree by following decision rules and reaches a leaf node that assigns a class label. Here I imported DecisionTreeClassifier from sklearn and did the same as previously.

	Precision	Recall	F1-score
0 (Non-churn)	0.83	0.86	0.85
1 (Churn)	0.57	0.52	0.55
Accuracy			0.77

Table 6. Decision tree performance on customer churn

The Decision Tree model demonstrates a moderate overall accuracy of 77%, indicating a decent ability to predict customer churn. Specifically, the model performs well in identifying non-churners (class 0) with a precision of 0.83, recall of 0.86, and an F1-score of 0.85. This suggests that the model effectively recognizes customers who are likely to stay with the company. However, the model struggles to predict churners (class 1), with a lower precision of 0.57, recall of 0.52, and F1-score of 0.55. This imbalance indicates that the model misses a significant portion of customers who are likely to leave, which could limit its effectiveness in targeting customers at risk of churning.

While the Decision Tree performs comparably to Naive Bayes in terms of overall accuracy, its difficulty in correctly identifying churners suggests a need for model improvement or class balancing to enhance its predictive power for churn behavior. Additionally, I also added a feature importance plot for the decision tree to see how it differentiated from the logistic regression one.



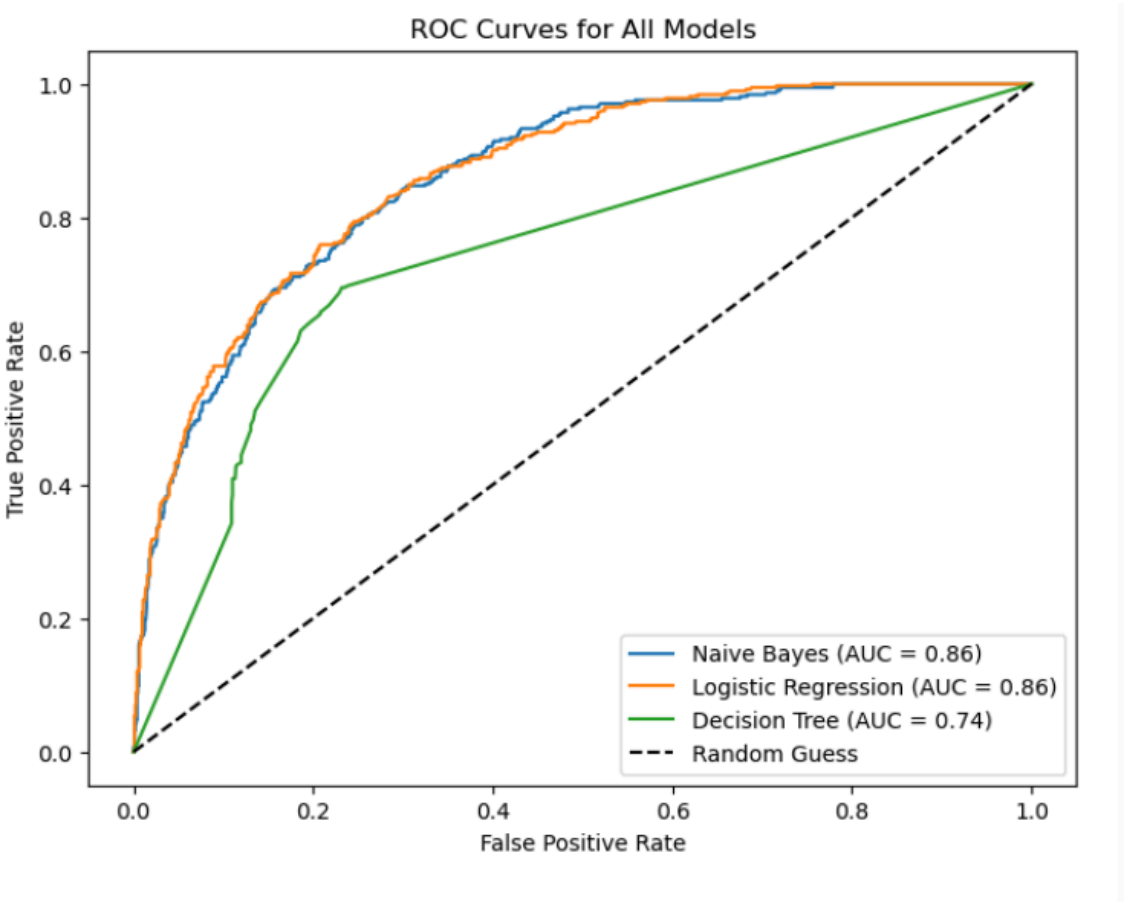
Graph 4. Feature Importance (Decision Tree)

The Decision Tree feature importance plot highlights **tenure** as the most influential factor in predicting churn, indicating that newer customers are more likely to leave. The **Month-to-Month contract** type is the second most important feature, supporting the idea that flexible contracts increase churn risk. Other significant features include the **number of services** (more services reduce churn) and being a **senior citizen** (associated with lower churn). In contrast, **internet type**, **area type**, and longer contract terms have minimal impact. This suggests that customer retention strategies should focus on new and month-to-month customers, according to the decision tree model.

5.4 Comparing the models

5.4.1 ROC and AUC

ROC (Receiver Operating Characteristic) Curve evaluates how well a model distinguishes between classes by plotting True Positive Rate (Recall) against False Positive Rate. AUC (Area Under the Curve) quantifies the ROC curve performance. If AUC is 1, it is a perfect classifier. If AUC is 0.5, it is just random guessing. To be able to evaluate the models even further, I chose to visualize the ROC curves and calculate the AUC scores. To do that I used the roc_curve and roc_auc_score functions from the sklearn.metrics module.



Graph 5. ROC Curves for All Models

Both the Naive Bayes and Logistic Regression models achieve an AUC of 0.86, indicating strong predictive capability in distinguishing between churners and non-churners. Their curves are close to the top-left corner, reflecting a high true positive rate and a low false positive rate. In contrast, the Decision Tree model has a lower AUC of 0.74, suggesting weaker performance and a greater likelihood of misclassification. Its curve is closer to the diagonal line, indicating it struggles more with separating churners from non-churners. Overall, the ROC analysis confirms that Naive Bayes and Logistic Regression outperform the Decision Tree, with Logistic Regression slightly more balanced and reliable for churn prediction.

5.4.2 Answering the RQs

For clarity, I have collected some of the results from the model testing in the table below.

Metric	Naive Bayes	Logistic Regression	Decision Tree
Accuracy	0.76	0.81	0.77
Precision (Churn)	0.54	0.65	0.57
Recall (Churn)	0.77	0.62	0.52
F1-score (Churn)	0.63	0.64	0.55
False Positives (FP)	246	124	146
False Negatives (FN)	86	142	178

RQ1: Which features are the most significant predictors of customer churn?

The analysis of feature importance across all models revealed that the most influential predictors of customer churn are the **contract type**, **tenure**, and the **number of services** a customer subscribes to. Specifically, customers on **month-to-month contracts** are significantly more likely to churn, whereas customers on **one-year** or **two-year contracts** show much lower churn rates. This aligns with the idea that longer contracts foster customer loyalty by creating commitment. Additionally, customers with **shorter tenure** are more prone to churn, suggesting that the early stages of the customer lifecycle are critical for retention efforts. Finally, customers subscribing to **fewer services** are more likely to leave, while those utilizing multiple services tend to stay, possibly due to the increased value they receive and the inconvenience of switching providers.

For businesses, these insights highlight actionable strategies to reduce churn. Companies should focus on encouraging customers to commit to longer contracts, perhaps through targeted promotions or discounts for annual plans. Additionally, enhancing the **onboarding experience** and delivering value early on can help retain new customers. Cross-selling or bundling services can also increase customer stickiness, as customers who engage with multiple services are less likely to switch to competitors. Overall, understanding these factors allows companies to design proactive retention strategies and improve customer lifetime value.

RQ2: How accurately can machine learning models predict whether a customer will churn?

The models demonstrated varying degrees of predictive accuracy. Logistic Regression outperformed the other models with an 81% accuracy and balanced performance across precision and recall. Naive Bayes followed with 76% accuracy, excelling in recall (77%) but struggling with precision (54%), leading to many false positives. Decision Tree had an accuracy of 77% but displayed weaker precision and recall, suggesting issues with overfitting and generalization.

Logistic regression is particularly effective for binary classification problems, like churn prediction, where the relationship between features and the target variable is relatively linear. It handles multicollinearity and categorical data efficiently, which is common in churn datasets. On the other hand, Naive Bayes, while strong in recall, assumes feature independence, which may not hold true for customer data. The Decision Tree's poorer performance was also anticipated due to its tendency to overfit on training data, especially with many features and small data variations.

RQ3: Which machine learning model performs best for predicting churn, and why?

Logistic Regression clearly outperformed both Naive Bayes and the Decision Tree across most metrics. Its balanced **precision (0.65)** and **recall (0.62)** for churn prediction, along with the highest overall **accuracy (81%)**, demonstrate its effectiveness. Furthermore, the **ROC-AUC score of 0.86** confirms its strong predictive capability. In comparison, Naive Bayes, while achieving the same AUC score, had a much lower precision and generated more false positives. Decision Trees lagged behind with an AUC of 0.74, reflecting weaker overall performance. Logistic Regression balances model complexity and predictive performance, avoiding overfitting while capturing key relationships in the data. This model's consistent performance across multiple evaluation metrics makes it the most reliable tool for predicting churn in this context.

6. Conclusion

6.1 Conclusion

This project aimed to identify the key factors influencing customer churn and evaluate the effectiveness of various machine learning models in predicting churn. Through statistical analysis and predictive modeling, several meaningful insights emerged. The analysis revealed that contract type, tenure, and the number of services subscribed significantly influence customer churn. Logistic Regression emerged as the most effective model for predicting churn, outperforming Naive Bayes and Decision Tree models in accuracy and balance between precision and recall. These findings offer actionable strategies for businesses to reduce churn, such as promoting long-term contracts, enhancing early customer engagement, and encouraging service bundling.

6.2 Potential pitfalls and limitations

Despite valuable insights, several limitations may have impacted the study's results. Data quality issues, such as missing values and the exclusion of the churn_reason dataset, may have limited predictive performance and introduced bias. The dataset's imbalance, with more non-churning customers, likely affected model accuracy, especially in identifying churners. Feature selection based primarily on feature importance and hypothesis testing might have excluded some influential variables. Additionally, Logistic Regression's linear assumptions, Naive Bayes' independence assumption, and the Decision Tree's susceptibility to overfitting posed model-specific challenges. The absence of cross-validation somewhat restricted the reliability of model evaluation, and external factors like market competition and seasonal trends were not considered, potentially omitting important churn drivers.

6.3 Future improvements

Several improvements could be done to improve this report. Balancing the dataset through resampling techniques or adjusting class weights could address class imbalance. Advanced feature and selection methods can help capture more nuanced patterns in customer behavior. Exploring other more complex models, such as random forest, gradient boosting or neural networks, could improve predictive performance. Implementing k-fold cross-validation would provide more robust model evaluations. Lastly, integrating external data sources, including market trends and customer service interactions, could offer a more comprehensive understanding of churn drivers.

In summary, this study provides essential insights into customer churn and demonstrates how data-driven approaches can guide effective retention strategies. Focusing on contract types, tenure, and service engagement allows businesses to target high-risk customers proactively. While Logistic Regression proved to be the most effective predictive model, future improvements in data handling and modeling could lead to even more accurate and impactful business decisions.