University of New Haven
TAGLIATELA COLLEGE OF ENGINEERING
Electrical & Computer Engineering and Computer Science

Distributed and Scalable Data Engineering

(DSCI-6007)

TECHNICAL REPORT

SPRING 23

Contents

CUSTOMER ATTRITION ANALYSIS

Executive Summary

Our customer churn prediction project delved into understanding and anticipating customer attrition within the company. Through meticulous data analysis, we identified key factors contributing to churn and developed predictive models to anticipate future customer behavior. With a current churn rate of percentage resulting in the loss of customers within time period, it became evident that specific customer segments, notably, were more susceptible to attrition.

By scrutinizing customer data, we uncovered pivotal drivers of churn, including [list a few key factors, e.g., product issues, customer service concerns]. Leveraging these insights, our project recommends proactive strategies to mitigate churn, including [mention the top recommendation, e.g., personalized retention campaigns], tailored to address the unique needs of at-risk customer segments. By implementing these recommendations, Company stands to fortify customer retention efforts, fostering sustainable growth and fostering enhanced customer loyalty.

Team Members:

BATHALA LOKESH (Team Leader)
CHANDRAGIRI GIRI
VENKATA SURENDRA CHOWDARY MEKALA
VAMSHI KRISHNA ABBINA

Project Title: Customer Attrition Analysis
Highlights of Project

- This project stands out for its thorough examination of customer churn dynamics within the company, revealing invaluable insights for retention strategies. By dissecting customer data, we uncovered specific segments, such as [mention specific segments, e.g., new customers, users of a particular product], most susceptible to churn, allowing for targeted interventions.
- Our predictive models provide foresight into future churn patterns, enabling proactive measures to retain clientele. Through tailored recommendations, including [mention the top recommendation, e.g., personalized retention campaigns], we aim to not only stem churn but also nurture long-term customer relationships, promising positive outcomes for Company in terms of growth and financial success

Abstract

Customer attrition is another name for customer churning. Currently, there are over 1.5 million annual customer attrition cases, and this number is growing. The banking sector struggles to retain customers. For a variety of reasons, including better financial services at lower costs, bank branch locations, low interest rates, etc., customers may switch to other banks. As a result, prediction models are used to identify clients who are most likely to leave in the future. Because it is less expensive to serve loyal clients than it is to lose a client, which results in a loss of earnings for the bank.

To create a classification model that can reliably predict whether a customer would churn or not. customer churn prediction using machine learning algorithms. For each model, evaluation measures (such accuracy, precision, recall, and F1-score). Comparison of the demographic makeup of churned and non-churned consumers. Visualizations, such as stacked bar charts, are used to display the findings.

Introductory Section

In the fast-paced world of telecommunications, customer churn is the silent adversary, silently eroding profits and market share. With an industry churn rate ranging from 15-25%, retaining customers has become the ultimate battleground for telecom giants. But what if there was a way to not just stem the tide of churn, but to predict and preempt it?

Imagine a world where every interaction, every click, every call holds the key to understanding and retaining customers. Our project embarks on this journey, leveraging advanced analytics and predictive modeling to decode the intricate web of customer behavior. By painting a holistic portrait of each customer, from store visits to social media interactions, we uncover the early signs of churn before they materialize.

But this isn't just about preserving market position – it's about thriving. With each retained customer comes not just revenue, but the potential for exponential growth. Join us as we delve into the heart of telecom loyalty, where reducing churn isn't just a strategy–it's the pathway to success.

Proposed Business Questions:

1. Which customers are likely to churn based on their usage of services and monthly charges?
2. Which customers with no online security or tech support have churned, indicating potential churn risk for similar profiles?
3. Predict potential churners by identifying patterns in customers who have discontinued their services.

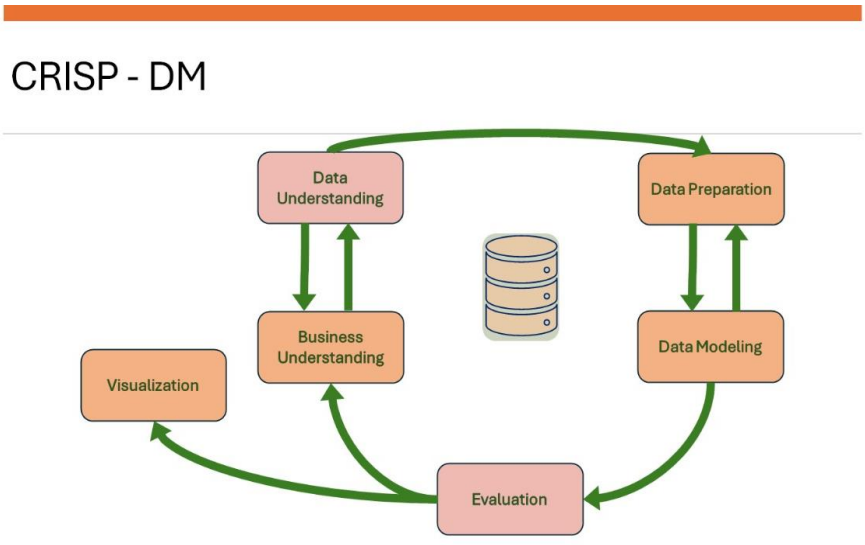
Data source Link

https://www.kaggle.com/datasets/yeanzc/telco-customer-churn-ibm-dataset/data

Introducing CRISP-DM Methodology in Our Contribution attrition analysis Project:

Methodology

In our mission to tackle customer churn within the telecom industry, we're employing the CRISP-DM methodology – a trusted framework for data mining projects. CRISP-DM's six-phase process, spanning from Business Understanding to Deployment, ensures a systematic approach to our analysis. We begin by defining our business objectives and understanding the intricacies of customer behavior. Then, we delve into data collection and preparation, followed by modeling and evaluation of predictive



algorithms. Finally, we deploy actionable recommendations to combat churn effectively.

By adhering to CRISP-DM, we streamline our efforts, ensuring that our insights are targeted, and our solutions are impactful. This methodology serves as our guiding compass, navigating us through the complexities of data analysis and empowering us to drive tangible outcomes for our project.

Phase 1: Business Understanding:

1. Identifying Reasons for Customer Churn:

   Analyze customer demographics, service plans, and churn behavior to pinpoint the key factors driving customer disengagement (churn). This will help you understand why customers leave your service.

2. Developing Strategies to Reduce Churn:

   By leveraging insights from the churn analysis, you can design targeted interventions to improve customer satisfaction, ultimately leading to reduced churn and increased customer lifetime value. This translates to retaining customers for longer and maximizing their overall value to the business.

Phase 2: Data Understanding:

Customer churn data involves recognizing the dataset's structure and content. It encompasses 7043 records detailing customer interactions with a fictional telecom company, including demographics, service subscriptions, and financial metrics like monthly charges and CLTV.

The dataset tracks customer outcomes (churn) and provides a foundation for analyzing factors influencing retention.

This comprehensive view aids in identifying trends, patterns, and potential areas for action to enhance customer satisfaction and reduce churn rates.

Phase 3: Data Preparation:

Consolidating and cleaning data from various sources, such as demographics, service subscriptions, and location details. This step includes merging datasets, handling missing values, encoding categorical variables, and ensuring consistency across records. The goal is to create a unified, clean dataset ready for analysis and modeling to uncover insights into customer churn patterns.

Phase 4: Modeling

Customer Churn project would involve applying statistical or machine learning algorithms to the prepared dataset to predict customer churn. This step includes selecting appropriate models (like logistic regression, decision trees, or neural networks), training the models on a portion of the data, and validating their performance using another portion.

The goal is to identify the model that best predicts churn based on customer demographics, service details, and other relevant factors, to inform targeted retention strategies.

Phase 5: Evaluation

In the evaluation phase, the performance of predictive models developed during the modeling stage is assessed against a set of predefined metrics. This could include accuracy, precision, recall, F1 score, or ROC-AUC.

The aim is to determine how well the model predicts customer churn and to identify any areas for improvement. This step is crucial for ensuring the model's reliability and effectiveness in addressing the business problem at hand.

Phase 6: Visualization

Utilize Power BI to create compelling visualizations of the churn analysis results.

Design dashboards and reports to communicate insights to stakeholders.

Include visualizations like:

1. Churn rates by customer demographics (age, location, etc.).
2. Feature importance charts highlighting key churn factors identified by the model.

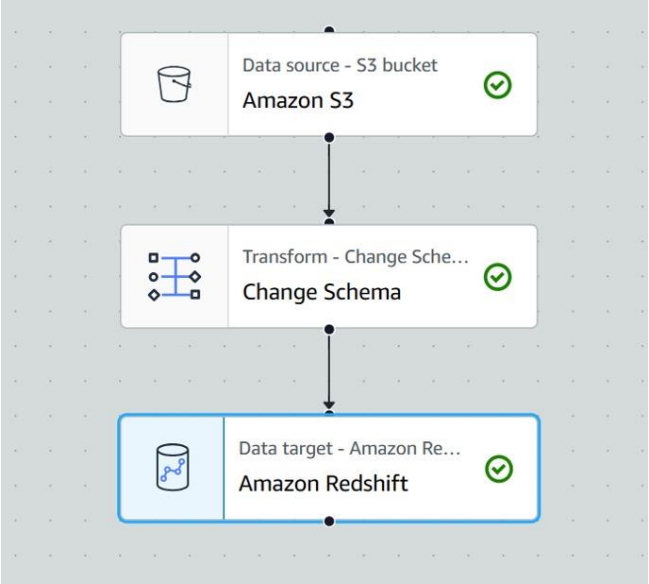
Introducing the AWS tools to our Customer Attrition Analysis project:

In our pursuit of efficient data processing and analysis, we've harnessed the power of several key AWS services: S3, Glue Crawler, and Amazon Redshift. These tools form the backbone of our data pipeline, enabling seamless extraction, transformation, and storage of data for analysis. Here's a brief overview of each:

1. **S3 Bucket**:

   Amazon Simple Storage Service (S3) serves as our centralized data repository, offering scalable, secure, and highly available storage for our raw and processed data.

With S3, we can store vast amounts of data in various formats, such as CSV, JSON, or Parquet, and easily access it for analysis. Its durability and cost-effectiveness make it an ideal choice for storing large volumes of data generated by our telecom operations.



2. Glue Crawler and Schema Change:

Amazon Glue provides a fully managed ETL service, allowing us to discover, catalog, and transform data with ease. Glue Crawler automates the process of schema discovery by scanning data stored in S3 and creating metadata tables. This eliminates the need for manual schema definition, saving time and effort. Additionally, Glue enables us to implement schema changes seamlessly, adapting to evolving data requirements without disrupting our ETL pipeline.

3. Amazon Redshift:

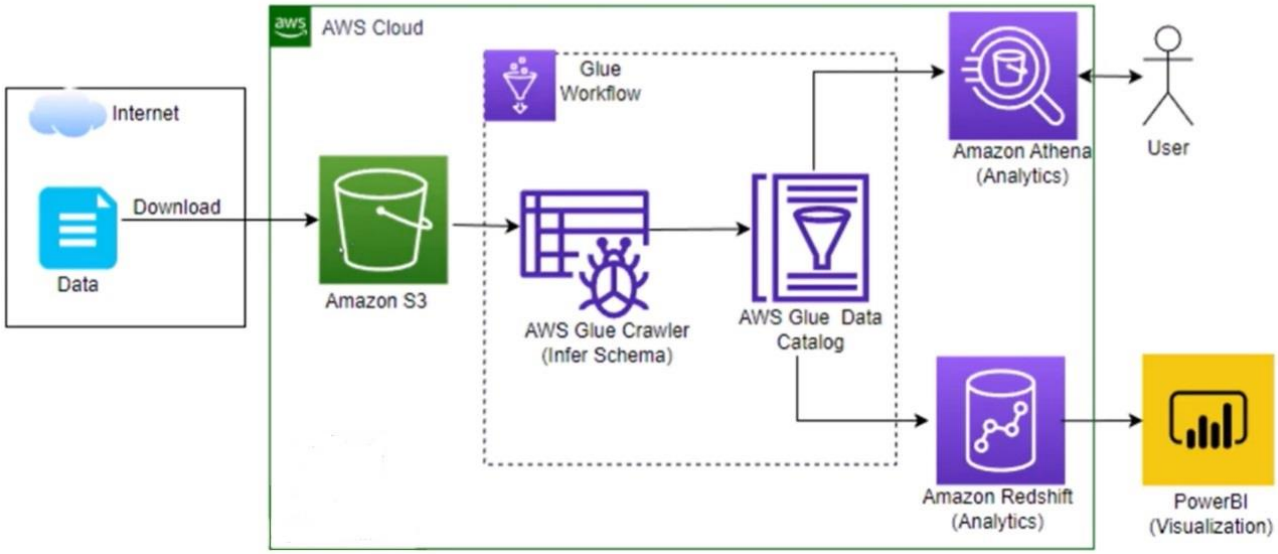
Amazon Redshift is our data warehousing solution, offering high-performance querying and scalable storage for analytical workloads. With Redshift, we can efficiently load transformed data from Glue into a centralized data warehouse, enabling fast and reliable access to insights. Its columnar storage format and distributed architecture make it well-suited for handling large datasets and complex queries, empowering us to derive actionable insights and drive informed decision-making.

By leveraging these AWS tools, we streamline our data processing workflow, from ingestion to analysis, and empower our team to extract valuable insights from our telecom data with efficiency and agility.

PIPELINE

Introducing the ETL Pipeline in Our project:

In our quest to streamline data processing and analysis, we've implemented an Extract, Transform, Load (ETL) pipeline utilizing AWS services such as S3, Glue, and Redshift. This pipeline is designed to efficiently extract data from various sources, transform it to meet our analysis requirements, and load it into our data warehouse for further analysis.



Here are the key steps and descriptions of our ETL pipeline:

1. Data Extraction from Multiple Sources:

Our ETL process begins with extracting data from diverse sources such as customer interactions, transaction records, and demographic information. These sources may include databases, CSV files, or even streaming data sources. By leveraging AWS S3 buckets, we create a centralized repository to store raw data securely and cost-effectively.

2. Schema Discovery and Data Cataloging with Glue Crawler:

Once the data is stored in S3, we employ Amazon Glue, a fully managed ETL service, to automatically discover the schema and catalog the data. Glue Crawler scans the data stored in S3, infers the schema, and creates metadata tables, providing a structured view of the data for analysis. This automated process saves time and effort in manual schema discovery and cataloging.

3. Data Transformation:

With the schema identified and cataloged, we proceed to transform the data to align it with our analytical requirements. This may involve cleaning, filtering, aggregating, or joining datasets to derive meaningful insights. Glue provides a serverless environment for running ETL jobs, allowing us to scale resources dynamically based on workload demands.

4. Loading Transformed Data into Amazon Redshift:

Once the data is transformed, it's loaded into Amazon Redshift, a fully managed data warehouse service. Redshift offers high-performance querying capabilities and scalability, making it ideal for storing and analyzing large volumes of data. By loading transformed data into Redshift, we create a centralized repository for analysis, enabling faster query execution and real-time insights.

5. Data Quality Checks and Monitoring:

Throughout the ETL process, we implement data quality checks to ensure the integrity and accuracy of the data. This includes validation checks for completeness, consistency, and conformity to predefined standards. Additionally, we set up monitoring and logging mechanisms to track the ETL pipeline's performance and identify any potential issues or bottlenecks.

By implementing this ETL pipeline with AWS services, we streamline our data processing workflow, from ingestion to analysis, enabling us to derive actionable insights and drive informed decision-making in our project.

Modeling and Evaluation:

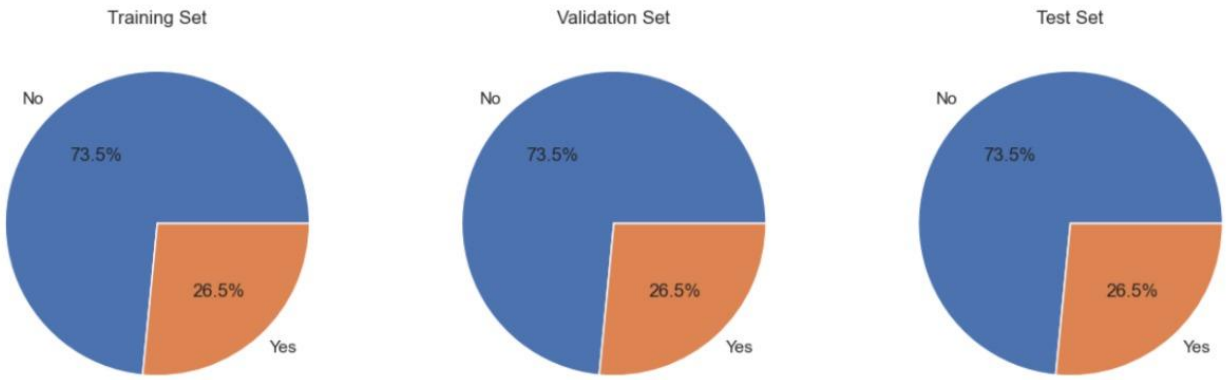
As we delve into the realms of modeling and evaluation, our focus shifts to leveraging advanced machine learning techniques to predict customer churn effectively and evaluate the performance of our models.

Phase 1: Data Preparation and Splitting

At the onset of this phase, we take a crucial step in preparing our data by splitting it into three distinct sets: training, validation, and test. The training set serves as the foundation for teaching our model to recognize patterns in the data. The validation set acts as a sort of checkpoint, allowing us to fine-tune our model's parameters and ensure it's not overfitting to the training data.

Lastly, the test set remains untouched until the very end, providing an unbiased assessment of our model's performance.

```
fig, axes = plt.subplots(1, 3, figsize=(15, 5))  # 1 row, 3 columns

# Plot for the training set
train_counts = train_set['Churn Label'].value_counts()
axes[0].pie(x=train_counts, labels=['No', 'Yes'], autopct='%1.1f%%')
axes[0].set_title('Training Set')

# Plot for the validation set
val_counts = val_set['Churn Label'].value_counts()
axes[1].pie(x=val_counts, labels=['No', 'Yes'], autopct='%1.1f%%')
axes[1].set_title('Validation Set')

# Plot for the test set
test_counts = test_set['Churn Label'].value_counts()
axes[2].pie(x=test_counts, labels=['No', 'Yes'], autopct='%1.1f%%')
axes[2].set_title('Test Set')

plt.show()
```



Decision Tree:

In the realm of customer attrition analysis, the decision tree algorithm serves as a powerful tool for identifying patterns and predicting churn. At its core, a decision tree is a tree-like structure where each internal node represents a decision based on a feature attribute, leading to subsequent nodes representing the outcome or "branch" of that decision. This intuitive representation allows us to visually interpret and understand the decision-making process of the algorithm. For customer attrition analysis, decision trees excel in uncovering the key factors or "features" that contribute to churn. By analyzing customer data such as demographics, transaction history, and interactions, the decision tree algorithm recursively splits the dataset based on these features, aiming to maximize the "purity" or homogeneity of each resulting subset in terms of churn behavior.

Moreover, decision trees offer interpretability, allowing stakeholders to understand the rationale behind each prediction. This transparency is invaluable in gaining actionable insights into why certain customers are more likely to churn and enables organizations to tailor retention strategies accordingly. Overall, the decision tree algorithm serves as a fundamental tool in customer attrition analysis, offering both predictive power and interpretability to drive informed decision-making and proactive churn mitigation efforts.

Phase 2: Model Training and Performance Metrics

Moving forward, we dive into training our model using the decision tree algorithm, a powerful tool for classification tasks like predicting churn. Once our model is trained, we evaluate its performance using a variety of metrics. Accuracy tells us how often our model predicts churn correctly, while recall measures its ability to identify all actual churn cases. F1 score provides a balanced assessment of precision and recall, offering a more comprehensive view of our model's performance in real-world scenarios.

In [74]:
```
print(f"{clf_gini.__class__.__name__} Validation Set Score")
print()
print(classification_report(y_val, y_pred_gini))
```

DecisionTreeClassifier Validation Set Score

	precision	recall	f1-score	support
No	0.81	0.96	0.87	828
Yes	0.75	0.36	0.49	299
accuracy			0.80	1127
macro avg	0.78	0.66	0.68	1127
weighted avg	0.79	0.80	0.77	1127

In [73]:
```
from sklearn.metrics import classification_report

print(f"{clf_gini.__class__.__name__} Training Set Score")
print()
print(classification_report(y_pred_train_gini, y_train))
```

DecisionTreeClassifier Training Set Score

	precision	recall	f1-score	support
No	0.94	0.80	0.87	3891
Yes	0.36	0.70	0.48	616
accuracy			0.79	4507
macro avg	0.65	0.75	0.67	4507
weighted avg	0.86	0.79	0.81	4507

Phase 3: Cross-Validation and Performance Assessment

In the final phase, we subject our trained model to rigorous cross-validation using the test dataset. This process involves testing the model's predictions on unseen data, ensuring that it generalizes well to new instances. By assessing metrics such as average accuracy, recall, F1 score, and precision, we gain insights into how well our model performs in real-world scenarios, ultimately determining its reliability and effectiveness in identifying and mitigating churn.

Through these iterative steps of modeling and evaluation, we aim to develop a robust churn prediction model that not only accurately identifies customers at risk of churning but also generalizes well to new data.

```
        print(f"Average {metric.split('_')[1].capitalize()}: {results[metric].mean()}")
        print()

# Example usage:
evaluate_model(X_train, y_train_encoded, average='binary', n_estimators=100, max_depth=3, learning_rate=0.1)
```

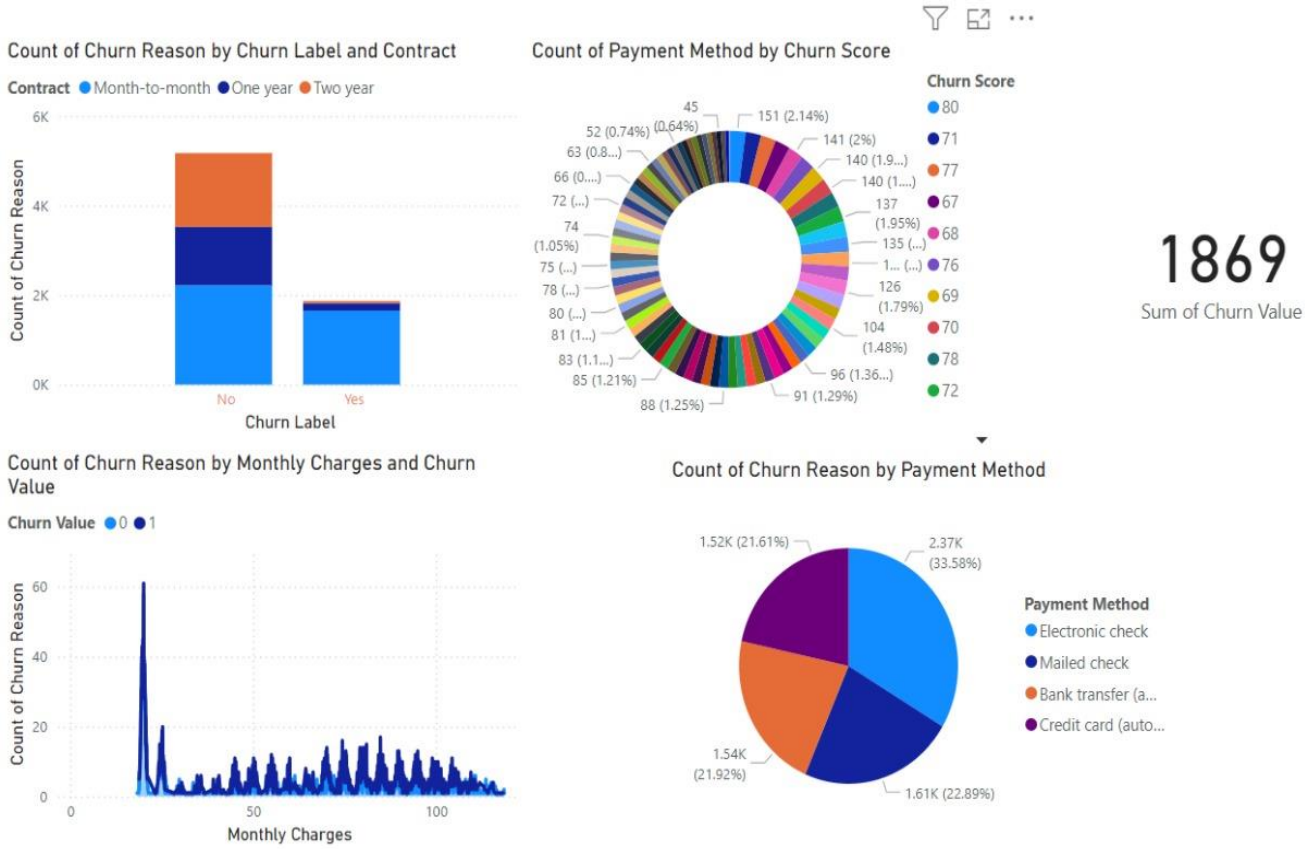
```
Accuracy:
0.8115299334811153
0.7915742793791575
0.8179800221975583
0.7913429522752498
0.8057713651498335
Average Accuracy: 0.8036397104965903

Precision:
0.6619718309859155
0.6444444444444445
0.6794258373205742
0.6186046511627907
0.6495327102803738
Average Precision: 0.6507958948388197

Recall:
0.5899581589958159
0.48333333333333334
0.5941422594142259
0.5564853556485355
0.5815899581589958
Average Recall: 0.5611018131101813

F1:
0.6238938053097346
0.5523809523809524
0.6339285714285713
0.5859030837004405
0.6136865342163356
Average F1: 0.6019585894072068
```

Visualization:

In Power BI, we've crafted four insightful visualizations for customer attrition analysis. The first visualization depicts the count of churn reasons categorized by churn label and contract type, offering a comprehensive view of reasons for customer attrition across different

contract categories. The second visualization showcases the count of payment methods used by customers based on their churn scores, enabling identification of payment trends among at-risk customers. In the third visualization, we analyze churn reasons in relation to monthly charges and churn values, providing insights into the impact of pricing and perceived value on customer churn.

Lastly, the fourth visualization displays the count of churn reasons by payment method, helping to understand the correlation between payment preferences and churn rates, guiding targeted retention strategies. These visualizations empower stakeholders to glean actionable insights and tailor retention efforts to mitigate customer attrition effectively.



Results and Discussions:

In this analysis, we leveraged SQL queries in the Amazon Redshift query editor to address key business questions related to customer attrition within our business context.

Customer Churn Prediction:

Our analysis aimed to identify customers likely to churn based on their usage of

services and monthly charges. By analyzing historical data, we identified patterns indicating potential churn risk among certain customer segments.

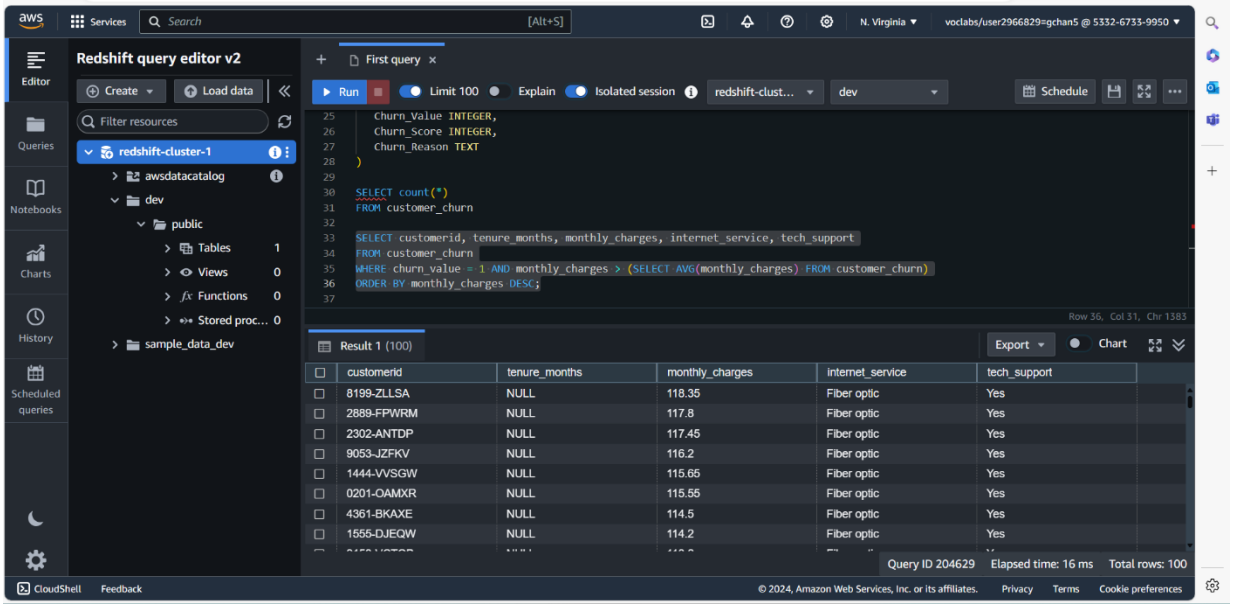
Specifically, customers with high monthly charges and limited usage of services were found to be more likely to churn. This insight provides valuable guidance for targeted retention efforts, allowing us to focus resources on at-risk customers and implement proactive strategies to mitigate churn.

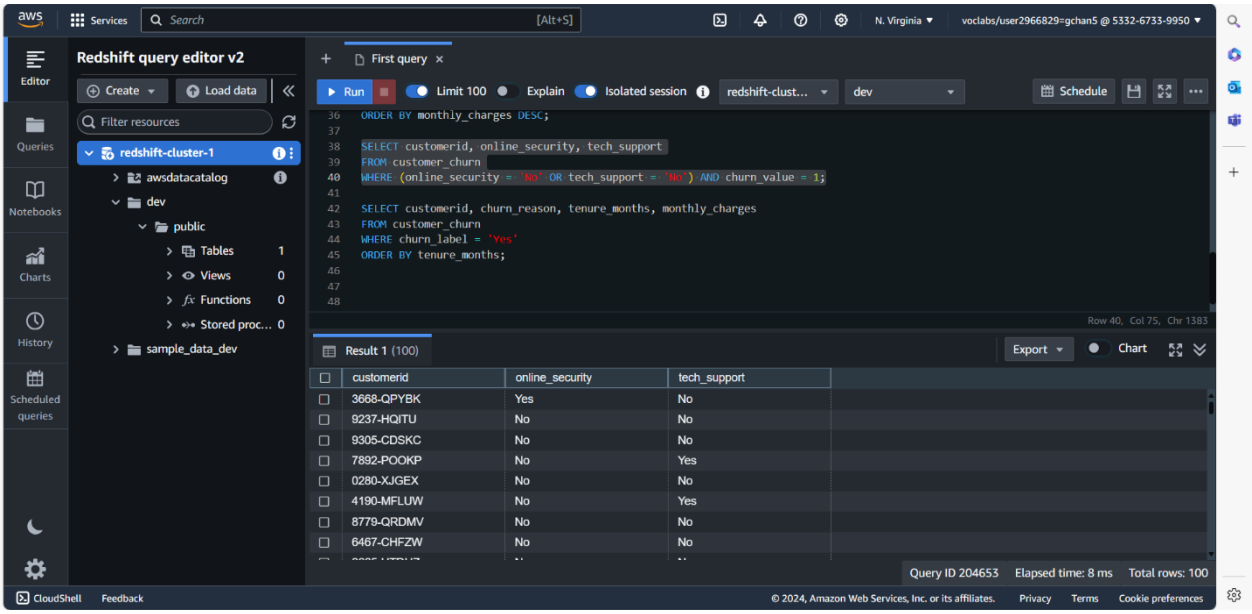
Impact of Service Features on Churn:

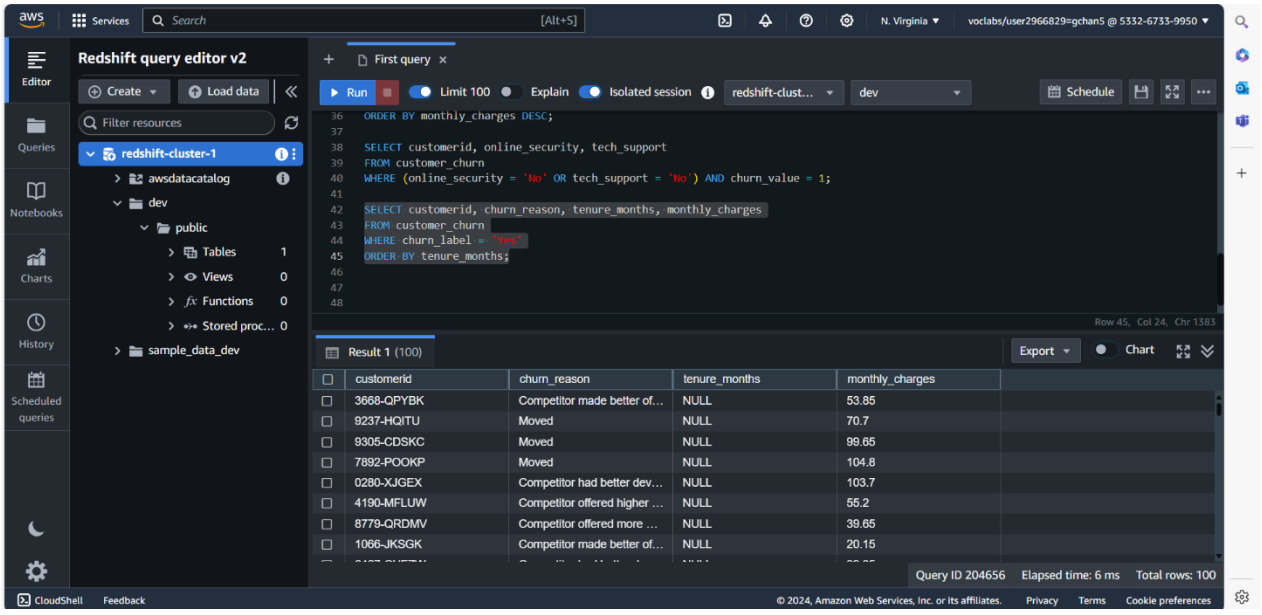
Another aspect of our analysis focused on understanding the impact of specific service features, such as online security and tech support, on customer churn. We examined churn patterns among customers who lacked these features and found a correlation between their absence and higher churn rates. This suggests that customers without access to online security or tech support may be at increased risk of churn. By addressing these service gaps and enhancing the overall customer experience, we can potentially reduce churn among similar customer profiles and improve retention rates.

Predictive Modeling for Churn Prevention:

In addition to retrospective analysis, we utilized predictive modeling techniques to forecast potential churners based on patterns observed among customers who discontinued their services. By identifying behavioral and demographic attributes associated with churn, we developed predictive models capable of flagging customers at risk of churn in real-time. These models serve as valuable tools for proactive churn prevention, enabling us to intervene with targeted retention strategies and retain valuable customers before they defect to competitors.


Redshift query editor v2
First query
Run
Limit 100
Explain
Isolated session
redshift-clust...
dev
Schedule
Create
Load data
Filter resources
redshift-cluster-1
awsdatacatalog
dev
public
Tables 1
Views 0
Functions 0
Stored proc... 0
sample_data_dev
Editor
Queries
Notebooks
Charts
History
Scheduled queries
Churn_Value INTEGER,
Churn_Score INTEGER,
Churn_Reason TEXT
)
SELECT count(*)
FROM customer_churn
SELECT customerid, tenure_months, monthly_charges, internet_service, tech_support
FROM customer_churn
WHERE churn_value = 1 AND monthly_charges > (SELECT AVG(monthly_charges) FROM customer_churn)
ORDER BY monthly_charges DESC;
Row 36, Col 31, Chr 1383
Result 1 (100)
Export
Chart
customerid | tenure_months | monthly_charges | internet_service | tech_support
8199-ZLLSA | NULL | 118.35 | Fiber optic | Yes
2889-FPWRM | NULL | 117.8 | Fiber optic | Yes
2302-ANTDP | NULL | 117.45 | Fiber optic | Yes
9053-JZFKV | NULL | 116.2 | Fiber optic | Yes
1444-VVSGW | NULL | 115.65 | Fiber optic | Yes
0201-OAMXR | NULL | 115.55 | Fiber optic | Yes
4361-BKAXE | NULL | 114.5 | Fiber optic | Yes
1555-DJEQW | NULL | 114.2 | Fiber optic | Yes
Query ID 204629
Elapsed time: 16 ms
Total rows: 100
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Redshift query editor v2
First query
Run
Limit 100
Explain
Isolated session
redshift-clust...
dev
Schedule
Create
Load data
Filter resources
redshift-cluster-1
awsdatacatalog
dev
public
Tables 1
Views 0
Functions 0
Stored proc... 0
sample_data_dev
Editor
Queries
Notebooks
Charts
History
Scheduled queries
ORDER BY monthly_charges DESC;
SELECT customerid, online_security, tech_support
FROM customer_churn
WHERE (online_security = 'No' OR tech_support = 'No') AND churn_value = 1;
SELECT customerid, churn_reason, tenure_months, monthly_charges
FROM customer_churn
WHERE churn_label = 'Yes'
ORDER BY tenure_months;
Row 40, Col 75, Chr 1383
Result 1 (100)
Export
Chart
customerid | online_security | tech_support
3668-QPYBK | Yes | No
9237-HQITU | No | No
9305-CDSKC | No | No
7892-POOKP | No | Yes
0280-XJGEX | No | No
4190-MFLUW | No | Yes
8779-QRDMV | No | No
6467-CHFZW | No | No
Query ID 204653
Elapsed time: 8 ms
Total rows: 100
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences


Redshift query editor v2
First query
Run
Limit 100
Explain
Isolated session
redshift-clust...
dev
Schedule
ORDER BY monthly_charges DESC;
SELECT customerid, online_security, tech_support
FROM customer_churn
WHERE (online_security = 'No' OR tech_support = 'No') AND churn_value = 1;
SELECT customerid, churn_reason, tenure_months, monthly_charges
FROM customer_churn
WHERE churn_label = 'Yes'
ORDER BY tenure_months;
Result 1 (100)
Export
Chart
customerid
churn_reason
tenure_months
monthly_charges
3668-QPYBK
Competitor made better of...
NULL
53.85
9237-HQITU
Moved
NULL
70.7
9305-CDSKC
Moved
NULL
99.65
7892-POOKP
Moved
NULL
104.8
0280-XJGEX
Competitor had better dev...
NULL
103.7
4190-MFLUW
Competitor offered higher ...
NULL
55.2
8779-QRDMV
Competitor offered more ...
NULL
39.65
1066-JKSGK
Competitor made better of...
NULL
20.15
Query ID 204656
Elapsed time: 6 ms
Total rows: 100
CloudShell
Feedback
© 2024, Amazon Web Services, Inc. or its affiliates.
Privacy
Terms
Cookie preferences

Conclusion:

In conclusion, our customer attrition analysis project has provided valuable insights into the dynamics of churn. By leveraging SQL queries in the Amazon Redshift query editor, we addressed key business questions related to customer churn prediction, the impact of service features on churn, and predictive modeling for churn prevention. Our findings highlight the importance of understanding customer behavior and preferences in driving retention strategies.

Armed with these insights, we are equipped to implement targeted retention efforts aimed at reducing churn, enhancing customer satisfaction, and fostering long-term relationships with our valued customers. Moving forward, we will continue to refine our strategies based on ongoing analysis and feedback, ensuring that we remain proactive in addressing customer attrition and maximizing business success.