Google Collab Link: Telco Customer Churn

1.0 Introduction

With the rise of artificial intelligence and the rapid advancements made by technologies today, data has become even more crucial in this data-driven world. Organizations today produce and collect enormous amounts of data. This makes data mining a vital tool in revealing hidden patterns and trends in these vast amount of data available. Companies from all sectors like healthcare to banking are adopting data mining techniques to transform raw data into actionable insights. For instance, economists in banks leverage data mining processes to predict market trends and other macroeconomic changes that often have vast amounts of data. In this case study, the author will investigate the use of data mining, looking at how it might be applied to better decision-making in actual case scenarios. This case study demonstrates how the data mining method can reveal hidden patterns, providing a deeper comprehension of the issue and assisting in the process of making better decisions.

1.1 Background of Study

The longevity of a business often depends on its ability to maintain a strong existing customer base and generate consistent demand for its product or service. Maintaining customers is key for long term success due to it being far more cost-effective than acquiring new ones (O'Brien and Downie, 2024) . In fact, it costs businesses three new customers for every one customer they lose due to high customer acquisition costs (O'Brien and Downie, 2024). This is where *customer churn* becomes a critical concept. *Customer churn* refers to the number of existing customers lost over a given period of time (O'Brien and Downie, 2024; Wagh *et al.*, 2024). It is a vital metric for understanding customer satisfaction and customer loyalty and allows businesses to identify potential red flags for revenue loss such as when churns are high. Subscription-based businesses, such as Software-as-a-Service (SaaS) platforms, telecom providers, and streaming services, highly benefits from understanding customer churn. This is because these businesses adopt a monthly recurring revenue (MRR) model and so sudden changes in customer numbers can cause immediate disruptions in financial forecast (O'Brien and Downie, 2024) . In the telecommunication (telcom) industry, there are millions of users who generate data daily and these companies often struggle

to identify why their customers are leaving, what behaviours indicate potential churn what can be done to prevent it (Wagh *et al.*, 2024). This is where data mining becomes especially useful. Data mining refers to the technique of examining big, complicated information to find significant patterns and connections that are not immediately apparent is known (Twin and Yashina, N. (2024). This case study aims to apply data mining techniques to a telcom churn dataset, that contains a variety of customer demographics, account lifecycles, usage behavior and service preferences, to identify the key drivers of churn and offer insights on the causes of churn and when to detect early signs of potential churn for businesses to take early action and so maintain their existing customers.

1.2 Problem Statement

The consequences of high churn poses a serious threat to a company's long-term success and calls for strategic ways for early detection and prevention. Despite having access to comprehensive customer data, many telcom providers struggle to pinpoint the root causes of churn and implement effective retention strategies. Three critical challenges persist:

1. The rapidly evolving nature of customer preferences and usage patterns, which makes behavioral prediction difficult.
2. The complex interplay between service-related attributes that directly influences customer decisions and drive brand loyalty.
3. The absence of a clear, data-driven framework to identify and proactively target high-risk customer segments.

To address these gaps, this study applies explainable, adaptable data mining techniques to balance accuracy with usability. By focusing on simpler yet robust methods, the author demonstrates how companies can detect churns earlier, understand *why* customers leave, adapt to behavioural shift and translate insights into targeted retention efforts.

1.3 Objectives

The objective of this study is as follows:

- To identify key customer attributes associated with churn by having comparison churn rates across demographic (e.g., gender, senior citizen), economic (e.g., monthly charges, payment method), and service-related features (e.g., contract type, internet service).
    - **Expected outcome:** a ranked list of customer attributes most associated with churn, supported by churn rate comparisons across demographics, payment methods, and service usage patterns.
- To determine how different combinations of service features most attributed to churn influence whether a customer churns (e.g., contract, tech support, internet service, online backup).
    - **Expected outcome:** clear identification of combinations of services that are associated with higher churn.
- To segment customers based on churn risk by grouping customers into distinct profiles using key features (e.g., tenure, contract type, and monthly charges) and identifying high-risk segments with significantly higher churn rates.
    - **Expected outcome:** defined customer segments with varying levels of churn risk, including a description of key characteristics of high-risk groups (e.g., short-tenure, month-to-month contracts, high monthly charges), to guide early intervention strategies.

2.0 Literature Review

2.1 Related Works

Recent research has demonstrated the value of combining fundamental data mining techniques for customer churn prediction. In the e-commerce domain, Xiahou and Harada (2022) employed a two-stage approach using clustering (k-medoids algorithm) to segment customers based on behavioral patterns, followed by classification (AdaBoost with decision tree weak learners) to predict churn within each cluster. Their preprocessing involved Min-Max scaling for normalization and one-hot encoding for categorical variables, achieving 82% accuracy on their 50,000-user dataset after addressing class imbalance with SMOTE.

Mitkees et al. (2017) applied a comprehensive data mining pipeline to telecom data, beginning with density-based clustering (DBSCAN algorithm) to identify natural groupings in customer behavior. They then implemented both classification (decision trees and multilayer perceptrons) and association rule mining (FP-Growth algorithm) to uncover patterns like "customers with 3+ billing complaints were 75% likely to churn." Their preprocessing included log transformations for skewed variables and label encoding for ordinal service types.

Another study by Mirjana *et al.* (2021) on churn management in telcom developed a hybrid telecom churn model combining k-means clustering (k=5, Euclidean distance) with decision tree classification. After standardizing 12 key features including monthly charges and service duration, they achieved 84% precision in identifying high-risk clusters. Notably, their decision trees revealed interpretable rules like "cluster_3 (long-tenure high-spenders) churned primarily after service outages."

Vallabhaneni et al. (2023) took a purely unsupervised approach, applying k-means (k=4, silhouette score=0.6) to segment 15,000 telecom customers without labeled churn data. By analyzing cluster centroids, they identified a critical segment with high inactive periods (>90 days) and low service usage that correlated with 68% eventual churn. The study emphasized manual feature engineering, including temporal binning of usage patterns into weekly intervals.

Across these studies, several technical patterns emerge: (1) clustering (typically k-means or DBSCAN) serves as a valuable preprocessing step before classification, (2) tree-based methods (decision trees, Random Forests) often provide the best balance of accuracy and interpretability, and (3) proper data preparation - including scaling, encoding, and imbalance correction - significantly impacts model performance regardless of the chosen algorithms.

2.2 **Overview of the Dataset**

The dataset that the author used in this case study is from Kaggle, which includes basic demographic details, services usages, payment method, contract length and churn information of individuals. These include factors such as gender, dependents, telco service purchase (i.e. Streaming TV, Phone Service, Device Protection etc), contract details (i.e. contract length), payment details and customer's churn. The purpose of this dataset to perform data mining techniques to identify the early signs of customer churn.

The dataset consists of 7043 rows (customers) and 21 columns (features). The explanation of each column are as follows:

1. **Customer ID:** the unique ID assigned to each customer, which is categorical data.
2. **Gender:** the gender of the customer, which is a categorical data.
3. **Senior Citizen:** indicates whether the customer is a senior citizen (1) or not (0), which is a categorical data.
4. **Partner:** indicates whether the customer has a partner (Yes) or not (No), which is a categorical data.
5. **Dependents:** indicates whether the customer has dependents (Yes) or not (No), which is a categorical data.
6. **Tenure:** the number of months the customer has stayed with the company, which is a numerical data.
7. **Phone Service:** indicates whether the customer has a phone service (Yes) or not (No), which is a categorical data.
8. **Multiple Lines:** indicates whether the customer has multiple lines (Yes) or not (No; No phonse service), which is a categorical data.
9. **Internet Service:** indicates which internet service the customer have (Fiber optic, DSL, Other), which is a categorical data.
10. **Online Security:** indicates whether the customer has online security (Yes) or not (No; No internet service), which is categorical data.

11. **Online Backup:** indicates whether the customer has online backup (Yes) or not (No; No internet service), which is a categorical data.
12. **Device Protection:** indicates whether the customer has device protection (Yes) or not (No; No internet service), which is a categorical data.
13. **Tech Support:** indicates whether the customer has tech support (Yes) or not (No; No internet service), which is a categorical data.
14. **Streaming TV:** indicates whether the customer has streaming TV (Yes) or not (No; No internet service), which is a categorical data.
15. **Streaming Movies:** indicates whether the customer has streaming movies (Yes) or not (No; No internet service), which is categorical data.
16. **Contract:** the contract term of the customer (month-to-month, one year, two year), which is a categorical data.
17. **Paperless Billing:** indicates whether the customer has paperless billing (Yes) or not (No), which is categorical data.
18. **Payment Method:** the customer's payment method (electronic check, mailed check, bank transfer, credit card), which is a categorical data.
19. **Monthly Charges:** the amount charged to the customer monthly, which is a numerical data.
20. **Total Charges:** the total amount charged to the customer, which is a numerical data.
21. **Churn:** indicates whether the customer churned (Yes) or not (No), which is a categorical data.

3.0 Methods

3.1 Pre-processing

Data pre-processing is described as the process of cleaning, transforming and preparing raw data into a usable format for data mining processing, machine learning or other data science related tasks. In this case study, the author uses data pre-processing techniques of label encoding and one-hot encoding to prepare the dataset for data mining.

3.1.1 Cleaning Data Types and Handling Missing Values

Before implementing these data pre-processing techniques, the author runs quick summary to check for correct data types displayed and if there any missing values.

```
Data types and missing values:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
dtypes: float64(1), int64(2), object(18)
memory usage: 1.1+ MB
None
```

Diagram 1 – Output: Summary of data types and missing values

The summary shows that monthly charges is recognised as an object (string) instead of a float (numeric). It is most likely stored as a string due to some missing or blank values. We will convert it to numeric using *pd.to_numeric* and set *errors='coerce'* to mark errors as NaN and then we will drop NaN values.

The feature 'customerID' is dropped as it provides no useful predictive value for this case study. The author runs a summary of the first fiver rows to check if feature 'customerID' is dropped.

```
Type of df before after customerID: <class 'pandas.core.frame.DataFrame'>
First 5 rows of df after dropping customerID:
```

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines	InternetService	OnlineSecurity	OnlineBackup	DeviceProtection
0	Female	0	Yes	No	1	No	No phone service	DSL	No	Yes	No
1	Male	0	No	No	34	Yes	No	DSL	Yes	No	Yes
2	Male	0	No	No	2	Yes	No	DSL	Yes	Yes	No
3	Male	0	No	No	45	No	No phone service	DSL	Yes	No	Yes
4	Female	0	No	No	2	Yes	No	Fiber optic	No	No	No

Diagram 2 – Output: Summary of first five rows of features after dropping feature 'customerID'

3.1.2 Label Encoding

Majority of the features in this dataset are categorical features, which makes it difficult later when applying data mining techniques to the dataset as most machine learning models cannot directly process categorial data like “Yes” and “No” and instead require the data to be numerical (Xiahou and Harada, 2022). To do this, we will encode these categorical features into numerical values using label encoding for further analysis and modelling purposes. Label encoding assigns a unique numerical label to each category in the feature which allows algorithms to process the data effectively.

3.1.3 One-Hot Encoding

Many of the features, especially service-related ones like *Internet Service*, *Device Protection*, and *Online Security*, are categorical variables with more than two possible values (e.g., “No internet service”, “DSL”, “Fiber optic”). These types of variables cannot be directly used in machine learning models, which typically require numerical input. To address this, we apply one-hot encoding, a pre-processing technique that converts each unique category into a new binary (0 or 1) column (Xiahou and Harada, 2022). For instance, the *Internet Service* feature is split into multiple columns such as InternetService_DSL, InternetService_Fiber optic, and InternetService_No, where each row has a “1” in the column corresponding to its category and “0” elsewhere. This allows the model to interpret the presence or absence of each category explicitly without assuming any ordinal relationship between them (Xiahou and Harada, 2022). This transformation is crucial for ensuring that our machine learning models interpret the categorical information correctly during training and prediction

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	PaperlessBilling	MonthlyCharges	TotalCharges	Churn
0	0	0	1	0	1	0	1	29.85	29.85	0
1	1	0	0	0	34	1	0	56.95	1889.50	0
2	1	0	0	0	2	1	1	53.85	108.15	1
3	1	0	0	0	45	0	0	42.30	1840.75	0
4	0	0	0	0	2	1	1	70.70	151.65	1

Diagram 3 – Output: Summary of first five rows of one-hot encoded categorical features

3.2 Data Mining Techniques

Based on the findings on related works, data mining techniques that was commonly brought up were tree-based methods such Decision Trees, Random Forests as they often provide the best balance of accuracy and interpretability. This case study will implement these data mining techniques to the dataset to achieve the objectives of the case study which is to identify the key drivers of churn and offer insights on the causes of churn.

3.2.1 Random Forest Classification Model

Random Forest Classification (RF) is a supervised machine learning technique that builds numerous decision trees using different subsets of the training data. The algorithm would then aggregate the results from each tree to determine the final output, reducing the likelihood of overfitting and boosting overall accuracy. The Random Forest Classification Model's capacity to manage huge datasets well while remaining resilient to noise and missing values is one of its main features (Ullah *et al.,* 2019).

In the context of this dataset, which focuses on predicting customer churn based on various customer and service attributes (such as monthly charges, contract type, and tenure group), RF is especially useful. This algorithm not only performs well with high-dimensional data but also allows us to rank the features based on how strongly they influence the prediction outcome. By training the Random Forest model on this dataset, we are able to gain insights into which factors are most associated with customer churn.

```
🔍 Top 10 Features Most Associated with Churn (Random Forest Importance):
TotalCharges                          0.204484
MonthlyCharges                        0.172099
Contract_Month-to-month               0.051095
tenure_group_Short                    0.033399
OnlineSecurity_No                     0.032262
gender                                0.031304
TechSupport_No                        0.031215
InternetService_Fiber optic           0.027933
PaymentMethod_Electronic check        0.027028
Partner                               0.026376
dtype: float64
```

Diagram 4 – Output: Top 10 Features Most Associated with Churn

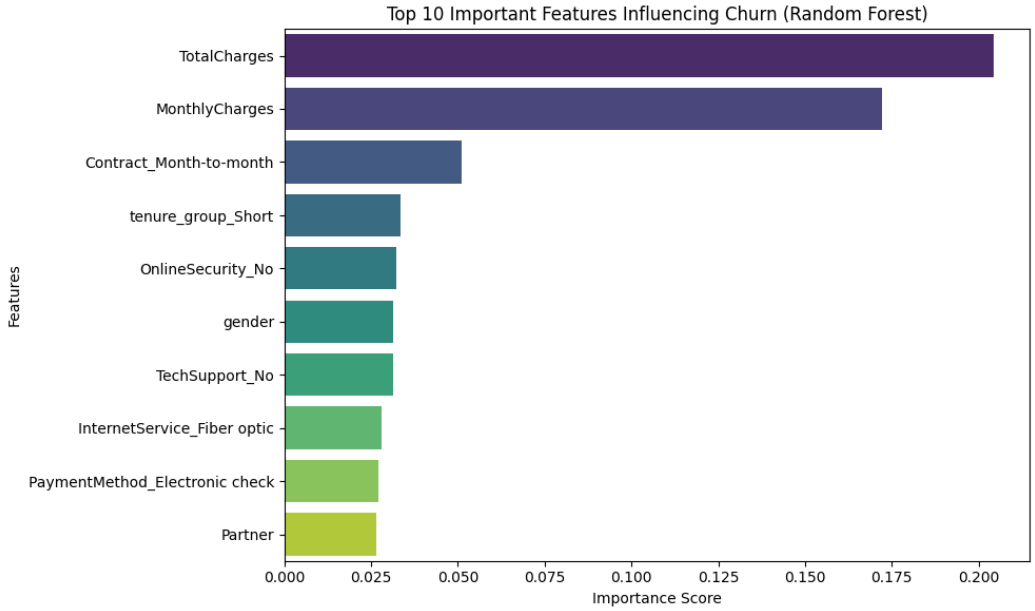


Diagram 5– Output: Random Forest Classification Model Visulisation

From the model's output, the most important predictor of churn is the Total Charges a customer has incurred, followed closely by Monthly Charges and whether the customer is on a Month-to-Month Contract. These findings provide a strong indication of which customers may be more at risk of leaving and help guide targeted retention strategy.

3.2.2 Decision Tree Classification Model

Decision Tree is a commonly-used data mining method that is ideal for examining consumer behavior due to its ease of use, transparency, and compatibility with both numerical and categorical data. Because business stakeholders can easily observe and grasp the decision flow, its unambiguous decision rules make it particularly helpful for churn analysis. This level of interpretability is important for companies looking to not only predict churn but also understand the "why" behind it (Mirjana *et al.*, 2021).

Another strength of Decision Trees is their ability to capture complex interactions between features without requiring any assumptions about data distribution (Mirjana *et al.*, 2021). For instance, in this case, the model picks up subtle but critical combinations of services that, when bundled together (or missing), significantly affect churn. This makes Decision Trees a practical and strategic tool to identify high-risk customer groups and tailor retention efforts accordingly.

In this dataset, we focus specifically on service-related features, such as internet service type, contract type, and support-related options, because these are the attributes that most directly influence a customer's experience and, by extension, their likelihood to remain loyal to the brand. These variables are operational touchpoints that businesses can actively improve, making them valuable for targeted churn prevention strategies.

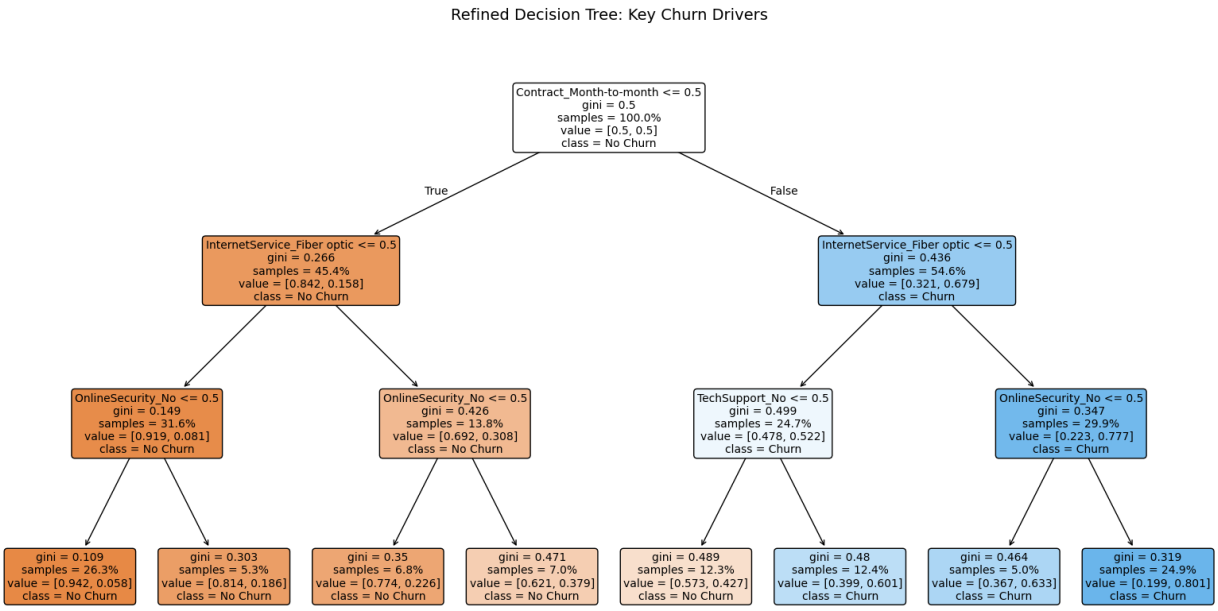
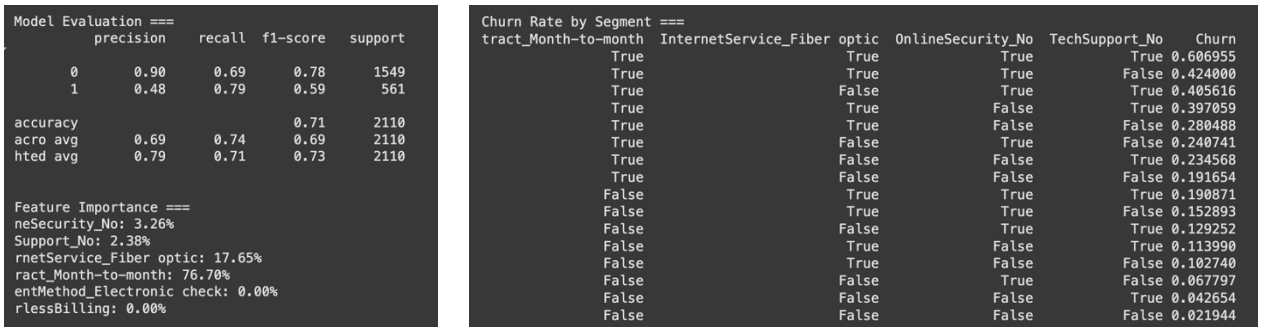


Diagram 6 – Output: Decision Tree Model Classification

To validate the insights derived from the Decision Tree model, a feature importance analysis and a segment analysis were conducted. Feature importance helps confirm which variables the model relied on most heavily when making predictions, giving us confidence that the tree is grounded in meaningful patterns (Mirjana *et al.*, 2021). Meanwhile, segment analysis helps us interpret how combinations of these key features behave in the actual dataset, offering a reality check against the model's predictions (Xiaohou and Harada, 2022). Together, these validation steps ensure that the decision rules derived from the tree are not only statistically sound but also practically relevant.



```
Model Evaluation ===
          precision    recall  f1-score   support

       0       0.90      0.69      0.78      1549
       1       0.48      0.79      0.59       561

accuracy                           0.71      2110
acro avg       0.69      0.74      0.69      2110
hted avg       0.79      0.71      0.73      2110


Feature Importance ===
neSecurity_No: 3.26%
Support_No: 2.38%
rnetService_Fiber optic: 17.65%
ract_Month-to-month: 76.70%
entMethod_Electronic check: 0.00%
rlessBilling: 0.00%
```

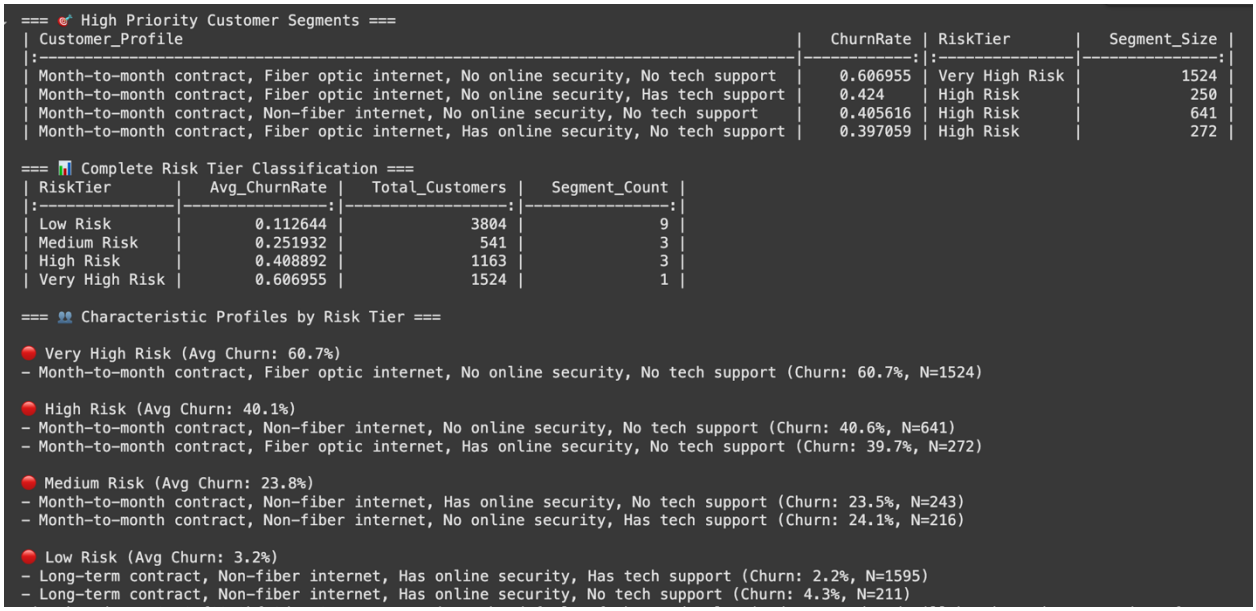
```
Churn Rate by Segment ===
tract_Month-to-month  InternetService_Fiber optic  OnlineSecurity_No  TechSupport_No    Churn
                True                         True               True            True 0.606955
                True                         True               True           False 0.424000
                True                        False               True            True 0.405616
                True                         True              False            True 0.397059
                True                         True              False           False 0.280488
                True                        False               True           False 0.240741
                True                        False              False            True 0.234568
                True                        False              False           False 0.191654
               False                         True               True            True 0.190871
               False                         True               True           False 0.152893
               False                        False               True            True 0.129252
               False                         True              False            True 0.113990
               False                         True              False           False 0.102740
               False                        False               True           False 0.067797
               False                        False              False            True 0.042654
               False                        False              False           False 0.021944
```

Diagram 6 – Feature Importance and Segment Analysis

By training a Decision Tree model using key service-related features, we find that Contract_Month-to-Month is the most important factor driving customer churn, followed by Internet Service: Fiber Optic. When further combined with customers who lack Online Security and Tech Support, the model uncovers a high-risk segment with a churn probability of 61%. This means that customers on month-to-month contracts who use fiber optic internet but do not have access to online security or technical support are significantly more likely to leave.

3.3 Customer Segmentation

Although Objective 3 aimed to segment customers by churn risk using key features, no advanced data mining or clustering techniques (e.g., k-means or decision trees) were applied, instead segments were created through basic groupings and manual rule-based classification.



=== 🎯 High Priority Customer Segments ===

Customer_Profile	ChurnRate	RiskTier	Segment_Size
Month-to-month contract, Fiber optic internet, No online security, No tech support	0.606955	Very High Risk	1524
Month-to-month contract, Fiber optic internet, No online security, Has tech support	0.424	High Risk	250
Month-to-month contract, Non-fiber internet, No online security, No tech support	0.405616	High Risk	641
Month-to-month contract, Fiber optic internet, Has online security, No tech support	0.397059	High Risk	272

=== 📊 Complete Risk Tier Classification ===

RiskTier	Avg_ChurnRate	Total_Customers	Segment_Count
Low Risk	0.112644	3804	9
Medium Risk	0.251932	541	3
High Risk	0.408892	1163	3
Very High Risk	0.606955	1524	1

=== 👥 Characteristic Profiles by Risk Tier ===

🔴 Very High Risk (Avg Churn: 60.7%)
- Month-to-month contract, Fiber optic internet, No online security, No tech support (Churn: 60.7%, N=1524)

🔴 High Risk (Avg Churn: 40.1%)
- Month-to-month contract, Non-fiber internet, No online security, No tech support (Churn: 40.6%, N=641)
- Month-to-month contract, Fiber optic internet, Has online security, No tech support (Churn: 39.7%, N=272)

🔴 Medium Risk (Avg Churn: 23.8%)
- Month-to-month contract, Non-fiber internet, Has online security, No tech support (Churn: 23.5%, N=243)
- Month-to-month contract, Non-fiber internet, No online security, Has tech support (Churn: 24.1%, N=216)

🔴 Low Risk (Avg Churn: 3.2%)
- Long-term contract, Non-fiber internet, Has online security, Has tech support (Churn: 2.2%, N=1595)
- Long-term contract, Non-fiber internet, Has online security, No tech support (Churn: 4.3%, N=211)

Diagram 7 – Customer Segmentation

4.0 Conclusion

This study shows how effective data mining methods can be in identifying and resolving client attrition. Through the methodical use of techniques like rule-based segmentation, decision tree modeling, and feature importance analysis, we were able to glean valuable insights from what would otherwise be unprocessed consumer data. These methods were essential for both determining the factors that contribute to churn and converting those discoveries into workable retention tactics.

From the start, proper data preprocessing laid the foundation for effective mining. Categorical features were encoded through both label and one-hot encoding to ensure compatibility with machine learning models while preserving the nuances between service options. This step was essential for transforming qualitative service data, such as internet type or contract terms—into a format that could be analyzed meaningfully, reinforcing the importance of robust data preparation in any analytical pipeline.

To address the first objective, data mining enabled a granular comparison of churn rates across different customer attributes. While initial exploratory analysis revealed some trends, the use of feature importance rankings through the Random Forest model allowed us to quantitatively validate which factors were most influential. Service-related attributes, particularly contract type, tenure, and monthly charges, consistently outperformed demographic factors in predictive power, highlighting the real drivers of churn.

For the second objective, the Decision Tree model was especially valuable. Its interpretable, rule-based structure enabled us to examine how specific combinations of features interact to influence churn likelihood. Unlike black-box models, the Decision Tree provided clear visibility into decision pathways, revealing that customers with month-to-month contracts, fiber internet, and no support services were most at risk. These insights could not have been as effectively uncovered through traditional analysis alone.

Addressing the third objective, we extended the insights from the Decision Tree model to conduct a manual rule-based segmentation, forming customer profiles based on their churn risk. While no complex clustering algorithm was used in this step, the segmentation logic was driven directly by the patterns surfaced through data mining. This resulted in a well-defined churn risk classification system, complete with detailed profiles that organizations can use to prioritize intervention. The segmentation output clearly differentiated between high-risk and low-risk customers, from a 60.7% churn rate for certain short-tenure, unsupported service combinations to 2.2% churn for long-term, fully supported customers.

Ultimately, this case study illustrates that data mining is not only a valuable analytical tool but also a practical one. It enabled us to move from raw customer records to strategic, insight-led decision-making. The techniques applied in this study—particularly feature importance scoring, interpretable decision tree modeling, and structured segmentation—provided clarity, direction, and evidence-based prioritization for churn reduction efforts.

Reference List

Mirjana, P.B, Pivar, J., Jaković, B. (2021) Churn Management in Telecommunications: Hybrid Approach Using Cluster Analysis and Decision Trees. *Journal of Risk and Financial Management (JRFM).* 14(11), pp. 544 – 569. [Accessed 9 May 2025].

Mitkees, I., Badr, S.M., ElSeddawy, A.I.B. (2017) *Customer churn prediction model using data mining techniques.* 2017 13th International Computer Engineering Conference (ICENCO), Giza, Egypt, 27-28 Decemeber. [Accessed 10 May 2025].

O'Brien, K and Downie, A. (2024) 'What is customer churn?', *IBM*, 9 September [online]. Available from: https://www.ibm.com/think/topics/customer-churn [Accessed 8 May 2025].

Twin, A. and Yashina, N. (2024) 'What is Data Mining? How It Works, Benefits, Techniques, and Examples', *Investopedia*, 23 February [online]. Available from: https://www.investopedia.com/terms/d/datamining.asp [Accessed 8 May 2025].

Ullah, I., Raza, B., Malik, A.K., Imran, M. (2019) A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector. *IEEE Access* [online]. 7, pp. 60134-60149. [Accessed 13 May 2025].

Wagh, S.K, Andhale, A.A, Wagh, K.S, Pansare, J.R, Ambadekar, S.P, Gawande, S.H (2024) Customer churn prediction in telcom sector using machine learning techniques. *Results in Control and Optimisation* [online]. 14(3). [Accessed 8 May 2025].

Xiaohou, X. and Harada, Y. (2022) B2C E-commerce Customer Churn Prediction Based on the K-Means and SVM. *Journal of Theoretical and Applied Electronic Commerce Research* [online]. 17(2), pp. 458-475. [Accessed 10 May 2025].

Individual Assignment Marking Rubrics

Criteria	Weightage	Excellent	Good	Satisfactory	Needs Improvement	Poor
Background of the Study	10%	Provides a comprehensive and well-contextualized background with clear relevance to data mining.	Provides relevant background with minor gaps in clarity or depth.	Basic context provided but lacks detail or specificity.	Vague or unclear background with limited relevance.	No meaningful background provided or completely off-topic.
Problem Statement	10%	Clearly defines a relevant and focused problem, well-aligned with data mining techniques.	Defines the problem adequately, with minor issues in clarity or alignment.	General problem stated but lacks specificity or strong connection to data mining.	Vague or weakly defined problem; relevance is not clear.	No clear problem stated or completely unrelated to data mining.
Objectives	10%	Objectives are specific, measurable, and directly linked to the problem and case study goals.	Objectives are mostly clear and relevant, with some room for refinement.	Basic objectives stated but lack detail or clarity in expected outcomes.	Objectives are vague, broad, or not clearly linked to the problem.	No objectives stated or completely unrelated to the case.