1. Title & Team Members

Project Title: Telecommunication Customer Churn Prediction Using Data Mining and Predictive Models

Course: CSCE 5380 - Data Mining

Team No: Group 3

Team Members:

- Shiva Sayi Shesshank Mamidi – 11752768
- Abhiram Reddy Pudi – 11817072
- Ramanaji Boggarapu – 11718646
- Sai Kumar Ramisetti – 11822974

2. Abstract

The telecoms sector regards customer attrition / Churn as a critical issue that demands immediate attention. This study utilizes a thorough and meticulous data mining process to identify the primary factors contributing to client attrition and to develop a reliable predictive model for future attrition events. We used the Telco Customer Churn dataset to find high-risk customer groups by combining supervised classification methods like SVM and Logistic Regression with unsupervised clustering methods. We found a significant group of "High-Risk Newcomer" customers and figured out which service combinations are most likely to lead to attrition. A tailored logistic regression model emerges as the most effective predictor, precisely identifying clients at risk. The results show that data mining may help businesses right away by giving them suggestions for targeted retention campaigns that are based on data.

3. Introduction & Research Question

Keeping customers is an important strategy for every business in the competitive telecoms industry. In this case, the costs of getting a new customer are sometimes five times higher than the costs of keeping an old one. When a corporation can predict client cancellations, it may adopt proactive tactics to keep those consumers by implementing retention approaches like as discounts or better customer service. Taking action to recover after losing a client is a reactive strategy, while taking action before losing a client is a more cost-effective strategy.

A comprehensive approach to data analysis aims to uncover the fundamental reasons behind client churn, notwithstanding the predictive power of machine learning algorithms. We need to find out the client profile, service composition, and behavioral patterns that have the biggest effect on customer turnover. Instead of just making basic predictions, this project uses real-world research to find out what causes clients to leave.

Conventional categorization approaches are the main ways to anticipate client attrition. The Logistic Regression is easy to understand, which helps businesses understand what factors affect client turnover and how much they do. Support Vector Machines (SVMs) are great at finding complex, non-linear connections between data. A decision tree provides clear "rules" that can be turned into business logic decisions that can be put into action. This study incorporates contemporary models into a broader data mining framework, including operational models to discern patterns.

Research Question: The primary research goal of this study is to answer the following empirical questions:

1. What are the key behavioural patterns, service configurations, and customer segments that are most strongly associated with customer churn in the telecommunications sector?

The study utilized data mining techniques to identify many important indicators significantly associated with churn. The "High-Risk Newcomer" client subgroup was the most important discovery. This group had a short engagement time (an average of 13 months), higher monthly charges (an average of $75), and a very high churn rate of 47.08%. The primary service configurations contributing to churn include Month-to-Month contracts, with churn rates of 40%, and Fiber optic internet access, particularly in the absence of protective add-on services such as Online Security and Tech Support. Behavioral data showed that short client lifespans and high monthly fees are the main reasons why customers leave, whereas long-term contracts and long-term relationships are the main reasons why customers stay.

2. Can a predictive model be trained to reliably identify at-risk customers, and is its performance statistically superior to other common baseline models?

A prediction model was created that accurately pinpointed clients at danger. Statistical analysis showed that this model worked better than the baseline models. The improved Logistic Regression model always did better than the other models. The model did a great job at telling the difference between people who churned and those who didn't. It got an AUC score of 0.8408 on the test set. The Logistic Regression model proved to be the most successful and reliable predictor for proactive customer retention methods, as demonstrated by a paired t-test ($P \text{-value} = 0.0025$) in comparison to the next best alternative, Decision Tree.

4. Methods and Experimental Design

The study employed a systematic approach to predictive modeling, incorporating detailed experimental design, data preparation, and data mining to identify trends :

4.1 Database :

The research utilized the publicly accessible "Telcom Customer Churn" dataset sourced from Kaggle. The dataset currently comprises 7,043 customer entries and 20 pertinent attributes, following the removal of the extraneous customerID.

- The demographic characteristics of consumers encompass gender, marital status, number of dependents, and senior citizen status.
- The account details encompass the duration of the agreement, payment methods, total cost, monthly payment amount, and account information.
- Paid phone, data, and internet plans that include additional services such as online security and technical assistance. We aim to monitor client churn, defined as the departure of clients within the past 30 days.

4.2 Data Cleaning for Analysis :

A pretreatment pipeline was established to ensure consistency in the process. The primary steps included:

- Strategies for Addressing Missing Data: Eleven clients with no duration of service exhibited a blank TotalCharges variable. Assigning a value of 0 to these is logical.
- TotalCharges has been converted to a numerical data type due to modifications in data types. The Churn variable consisted of two values: 1 indicating "Yes" and 0 indicating "No."
- OneHotEncoder was employed to encode categorical data, while StandardScaler was utilized to standardize the numerical features, specifically tenure, MonthlyCharges, and TotalCharges. A ColumnTransformer was employed to scale and encode these attributes. The direct utilization of the pipeline during model training effectively mitigated data leakage.

4.3: Data Mining Techniques

The two main data mining methods employed to solve the original research question were K-Means clustering and association rule mining.

- K-Means clustering : Models Based on Kernels To uncover the underlying customer segmentation, clustering used scaled numerical features. If you're utilizing the Elbow Method, you should try to get the best results with three clusters (k=3). The created segments were sorted by average qualities, with an emphasis on assigning churn rates.

- Association rule mining : The Aprior algorithm was used by Association Rule Mining to figure out which types of service subscriptions and contracts are linked to customer attrition. The rules were made and tested with a minimum support of 0.05 and a lift of greater than 1.1. The only rules that were looked at were the ones that said "Churn=Yes."

4.4: The models employed in experimental design and predictive :

We chose three categorization models for this comparison: Logistic Regression, Decision Tree, and Support Vector Machine.

- Setup for the experiment: Twenty percent of the dataset was used for both training and testing. A stratified split was used to repair an unbalanced dataset so that both groups had the same percentage of churners.

- We always utilized a random state of 42 for all non-deterministic activities in the project, like data splits and model initialization, so that the results could be repeated.

- To fix the imbalance between the two classes, we changed all three models such that the class_weight='balanced' option was used for the non-churner class, which is the larger of the two.

- We used GridSearchCV with 5-fold cross-validation to discover the optimum hyperparameters for each model. Then we used Area Under the ROC Curve (AUC) to test them. This way, we may compare the models' best potential variants.

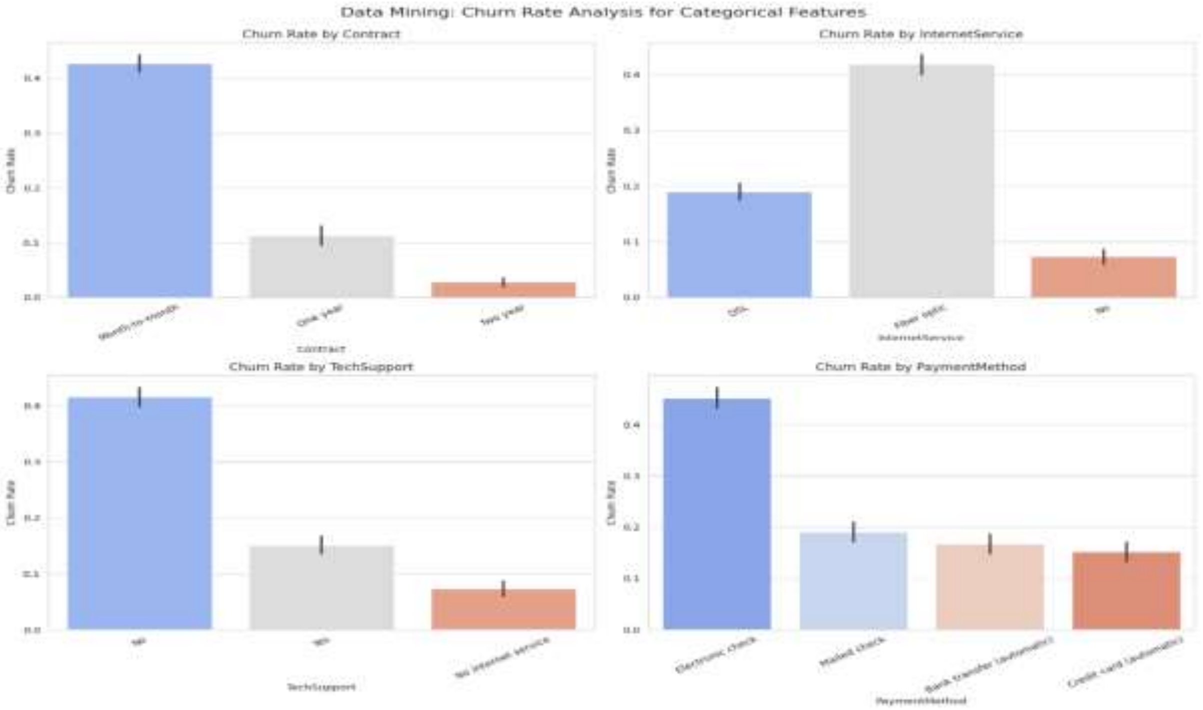
- Criteria for Evaluation: Accuracy, Precision, Recall, F1-Score, and AUC were some of the many ways used to examine how well the model did on the test set that it had never seen before. AUC was chosen as the main metric for comparing models because it is effective in situations where it is hard to categorize data and there is a big class imbalance.

- Quantitative Evaluation: We ran a paired t-test on the 10-fold cross-validation AUC values of the two best models to see if there was a statistically significant difference between them.

5. Results and Analysis

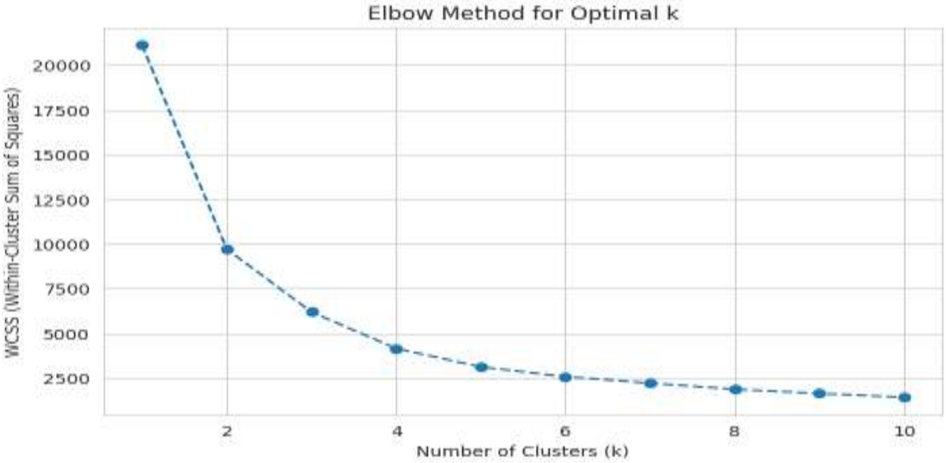
This section outlines the key qualitative and quantitative findings derived from our analysis using data mining and predictive modeling techniques.

5.1. Results of Data Analysis: Recognizing Trends in Customer Departure
Our exploratory analysis uncovered notable patterns. The data reveals that the overall turnover rate stands at 26.5%. This rate shows considerable variation across various client categories.



(Figure 1: Data Mining: Churn Rate Analysis for Categorical Features)

Figure 1 indicates that consumers with month-to-month contracts exhibit a turnover rate exceeding 40%, whereas those with two-year contracts demonstrate a churn rate below 5%. Clients with fiber optic internet are far more prone to disengagement than those with DSL. K-Means clustering facilitated our understanding by identifying three distinct client groups.

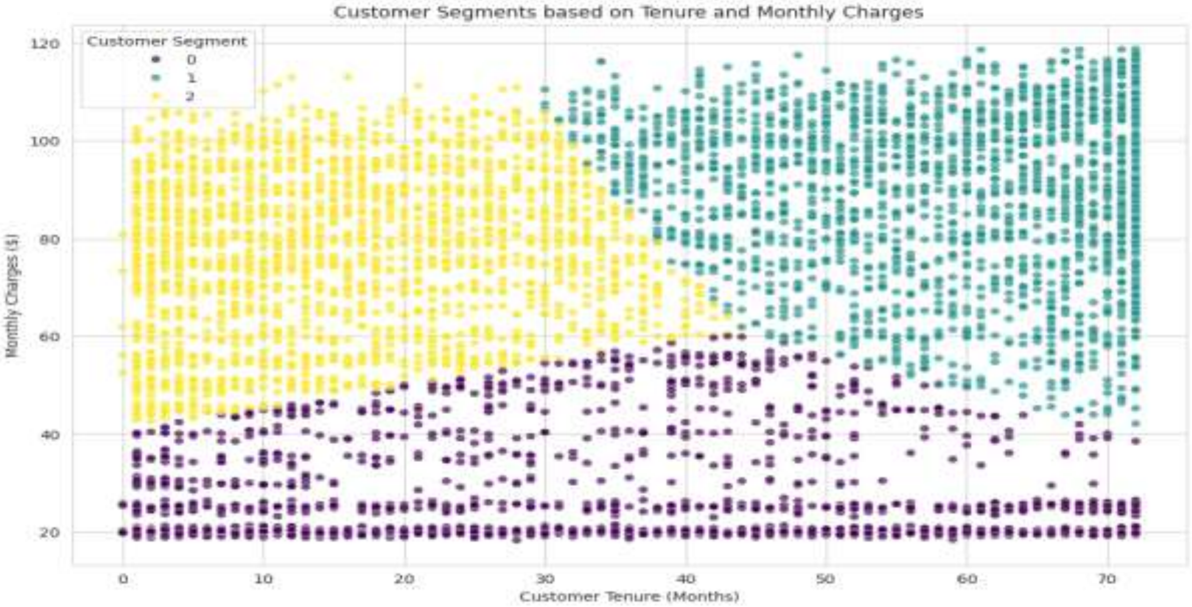


(Figure 2: Elbow Method for Optimal k)

Customer Segment	**SegmentSize**	**tenure**	**Monthly Charges**	**Total Charges**	**Churn Rate**
0	2154	29.50	26.57	809.42	12.30%
1	2200	58.56	89.70	5246.13	15.36%
2	2689	13.25	74.95	1030.57	47.08%

(Table 1: Characteristics and Churn Rate of Each Segment)

Table 1 illustrates the identification of three parts. The most intriguing discovery is segment 2, designated as the "High-Risk Newcomer" segment. This cohort consists of recent consumers that have been with the organization for an average duration of 13 months, contribute an average monthly payment of $75, and have a substantial attrition rate of 47%.



(Figure 3: Customer Segments based on Tenure and Monthly Charges)

Figure 3 shows that the clusters are separated. The high-risk Segment 2 (yellow) is in the area of the plot with low tenure and high monthly charges.

Lastly, Association Rule Mining found some pairings of services that were very likely to lead to churn.

Antecedents (Service Combinations)	Consequents	Support	Confidence	Lift
OnlineBackup=No, PaperlessBilling=Yes, Interne...	Churn_Yes	0.0574	0.7038	2.652
PaperlessBilling=Yes, InternetService=Fiber op...	Churn_Yes	0.0585	0.6913	2.605
OnlineBackup=No, PaperlessBilling=Yes, Interne...	Churn_Yes	0.0586	0.6883	2.594
OnlineBackup=No, PaperlessBilling=Yes, Interne...	Churn_Yes	0.0625	0.6864	2.587
OnlineBackup=No, PaperlessBilling=Yes, Interne...	Churn_Yes	0.0628	0.6842	2.578

(Table 2: Top 5 Rules Leading to Customer Churn)

The primary churn rules are displayed in Table 2. According to the most definite rule, 70% of fiber optic internet users will cancel their service if they do not have access to online backup or technical assistance. Here is a wealth of information that you may put to good use.

5.2. Predictive Modeling Results

The three models were tested on the held-out test set after being carefully tuned.

Model	**Accuracy**	**AUC**
Logistic Regression	0.738822	0.840755
Decision Tree	0.742370	0.836897
SVM	0.734564	0.836154

(Table 3: Summary of Final Model Performance on Test Set)

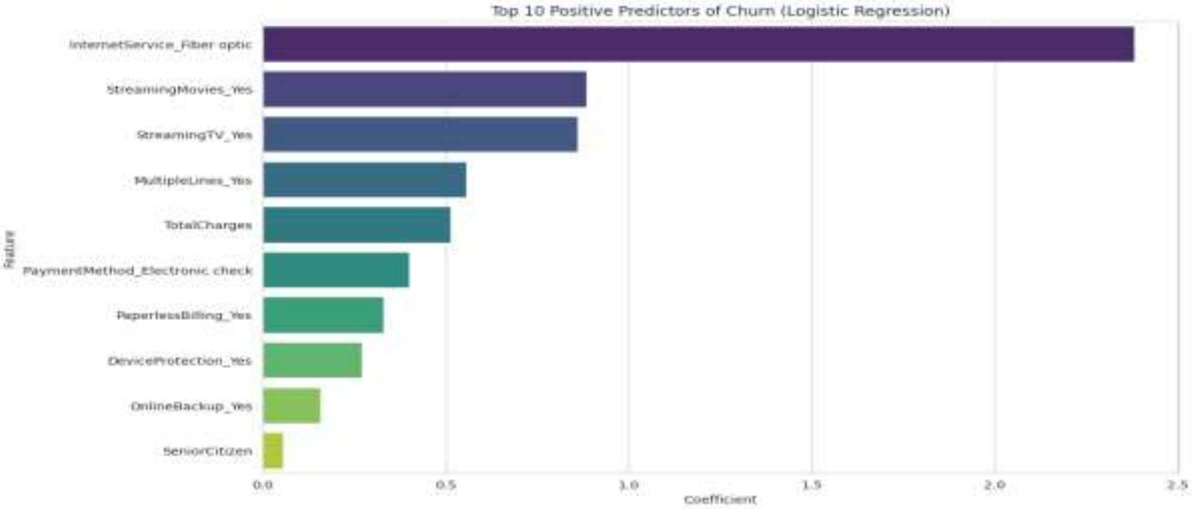
Table 3 shows peak efficacy metrics for each model. Logistic Regression had the highest AUC, 0.8408. This shows a better capacity to identify churners. The churn class model identified 78% of service discontinuers with a recall score of 0.78.

To validate this result, a paired t-test was conducted.

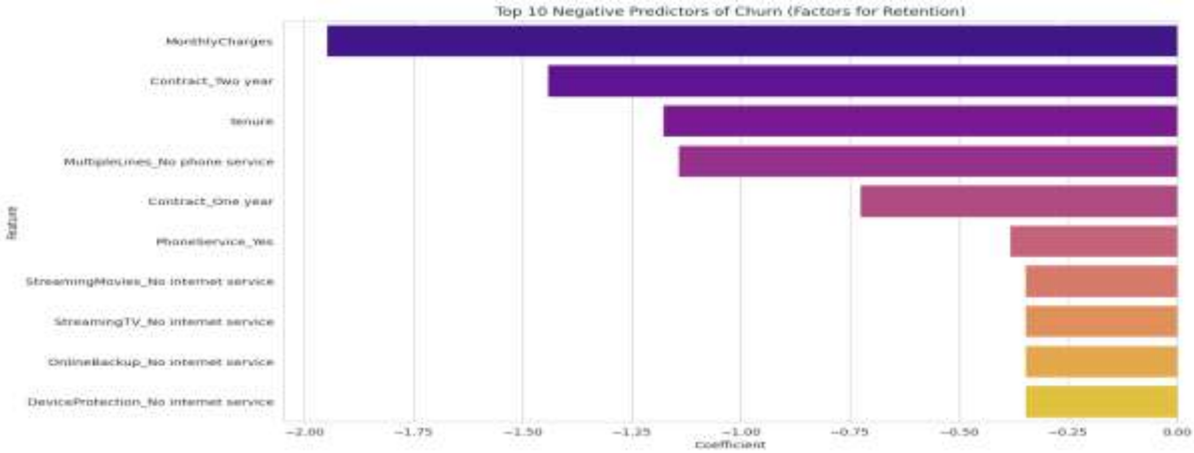
Statistic	**Value**
T-statistic	4.1433
P-value	0.0025

(Table 4: Paired t-test Results)

The results of the paired t-test for the two leading computer models—Decision Tree and Logistic Regression—are presented in Table 4. Given that the p-value of 0.0025 is markedly below the 0.05 threshold for statistical significance, we can infer that the enhanced performance of the Logistic Regression model is unlikely to be a mere coincidence.



(Figure 5: Top 10 Positive Predictors of Churn (Logistic Regression))



(Figure 6: Top 10 Negative Predictors of Churn (Factors for Retention))

Figures 5 and 6 display the picture-based coefficients obtained from the top Logistic Regression model. The most important factors influencing customer retention are length of tenure and Contract Two year, however InternetService Fiber optic is confirmed to be the primary predictor of churn.

6. Discussion and Interpretation

Our comprehensive investigation has yielded interrelated and illuminating insights concerning client attrition. The predictive models were employed to validate and quantify the forecasts and trends identified in the initial phase of the investigation. The evidence robustly corroborates our primary hypothesis that certain patterns unequivocally contribute to turnover.

The most significant announcement is the initiation of the "High-Risk Newcomer" segment. The 47% attrition percentage for this group indicates a significant issue with customer retention from the outset. This is particularly applicable to clients with expensive, non-binding plans. The forecasts of the algorithms could not be more accurate. A month-to-month contract and elevated monthly charges are both significant indicators of potential customer attrition. Fiber optic service is a significant indicator of turnover, which is unexpected given its status as a premium offering. This may indicate that the consumer perceives insufficient value or that the service lacks the anticipated reliability relative to its cost, perhaps leading to dissatisfaction among new users.

The strength of the findings is further confirmed by the consistency of the analytical methods employed. K-Means identified the high-risk group, Association Rules specified the hazardous service combinations within that group, and the predictive models (Decision Tree and Logistic Regression) all validated the notion that these parameters are essential for forecasting.

Constraints: The research presents several issues. The information is a static snapshot and cannot illustrate the evolution of a customer's trip over time. For instance, we would be unable to ascertain whether a customer's data usage or the frequency of support calls increased in the months preceding their departure. The dataset lacks information regarding customers' residences, which may significantly contribute to the confusion surrounding service quality charters.

7. Reflection on Methodology

Our data-driven technique exhibited numerous advantages. The application of both unsupervised (K-Means) and supervised (classification) methods not only led to greater insights but also resulted in better outcomes than employing any of the two techniques independently. Adoption of a potent experimental design consisting of GridSearchCV with cross-validation, class_weight='balanced' for dealing with the imbalance, and a concluding statistical t-test, has not only ensured that our conclusions are backed up by a single fortunate data split but also by scientific argumentation. Utilization of Pipelines and random_state guarantees that the whole workflow is reproducible.

Conversely, there are some methodological aspects where improvement is possible. Our solution to class imbalance was effective but too basic. Methods like SMOTE (Synthetic Minority Over-sampling Technique) might have been studied to see if the true positive rate, in particular, is improved by the model's performance through the production of fictional data for the under-represented class. Also, our feature engineering was limited to the basic features that we had. In addition, the interaction terms (e.g., tenure * MonthlyCharges) or more complex features could be created to capture the data's complex patterns. The choice of k=3 for clustering was made through the Elbow method, which is at times seen as subjective; thus, other cluster validation metrics could be used for support.

8. Conclusion & Future Work

Conclusion: This study employed a data-driven empirical approach to elucidate the comprehensive customer turnover process and its significant trends. An effort was made to identify a cohort of consumers labeled "Newcomers," who are highly predisposed to churn, and the modified logistic regression model emerged as the most effective statistical instrument for locating these customers. The main factors influencing client turnover are short-term contracts and Fiber Optic service, whereas long-term contracts and tenure are the key variables for customer retention.

Prospective Endeavors: Following this effort, alternative research avenues may be pursued:
Establish a survival analysis model: Rather than merely predicting whether a client will depart, consider employing a survival model (such as Cox Proportional Hazards) to estimate the timing of their potential exit. This aligns with retention tactics that are more focused and utilize less resources.

Incorporate Unstructured Data: You may include items such as transcripts of customer service calls or online reviews. We could employ Natural Language Processing (NLP) to ascertain individuals' sentiments and discussions. This would introduce highly potent predictive attributes on customer satisfaction that are currently unavailable.

Develop a Dynamic Churn Prediction System: A model may be developed regularly, or even more often, to analyze time-series data, such as monthly consumption and payment history, generating a real-time "churn risk score." This would provide a dynamic view of the customer's health rather than a static forecast.

Investigate more sophisticated models for interaction effects: Utilize gradient boosting models such as XGBoost or LightGBM. These models perform exceptionally with tabular data and can autonomously identify intricate correlations among characteristics. This may offer a modest yet significant enhancement in performance compared to Logistic Regression.

9. References

[1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2012.

[2] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit-driven data mining approach," *European Journal of Operational Research*, vol. 218, no. 1, pp. 211-229, Apr. 2012.

[3] K. Coussement and D. Van den Poel, "Churn prediction in subscription services: An application of support vector machines," *Expert Systems with Applications*, vol. 34, no. 1, pp. 489-498, Jan. 2008.

[4] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A churn prediction model using random forest: analysis of telecom industry," *IEEE Access*, vol. 7, pp. 60134-60149, 2019.

[5] H. Ahn, I. Han, and Y. Lee, "Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry," *Telecommunications Policy*, vol. 30, no. 10–11, pp. 552–568, Nov. 2006.

[6] "Telco Customer Churn," *Kaggle*, 2020. [Online]. Available: https://www.kaggle.com/datasets/blastchar/telco-customer-churn

[7] F. Pedregosa *et al*., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, Oct. 2011.

[8] W. McKinney, "Data structures for statistical computing in Python," in *Proc. 9th Python in Science Conf. (SciPy)*, 2010, pp. 51–56.

[9] S. Raschka, "MLxtend: Providing machine learning and data science utilities and extensions to Python’s scientific computing stack," *Journal of Open Source Software*, vol. 3, no. 24, p. 638, Apr. 2018.

10. Individual Contributions :

1. **Abhiram Reddy Pudi (11817072):**

Data Mining, Exploratory Analysis & Unsupervised Clustering: Oversaw the project's pattern discovery phase, mining the raw data for key insights and patterns before modeling.

Particular Contributions:

- To complete the Exploratory Data Analysis (EDA), we created and assessed churn distribution visualizations, analyzed churn rates across key categories, and compared churner and non-churner numerical data distributions.

- In unsupervised clustering, we ran the full gamut of K-Means clustering tests. In order to accomplish this, we used the Elbow Method to determine the optimal number of clusters, fitted the model, then profiled the resulting segments to identify the most significant "High-Risk Newcomer" category.

- Mining of Patterns: Employed Association Rule Mining technique to identify, with high certainty, the specific service-term combinations that cause customers to churn.

2. **Shiva Sayi Shesshank Mamidi (11752768):**

Predictive Modeling and Hyperparameter Tuning Role: Led the foundational predictive modeling phase, accountable for the systematic training and optimization of supervised learning models.

Particular Contributions:

- Model and Grid Definition This stage established the experiment to identify the most effective model. Model Selection: I selected three distinct classification algorithms—Logistic Regression, Decision Tree, and SVM—to facilitate a robust comparison of their underlying mechanisms, which include linear, rule-based, and non-linear boundaries. Hyperparameter Space: established a defined set of hyperparameters, including the strength of regularization and the depth of the tree, for evaluation in each model. The "grid" is significant as the model's performance is highly sensitive to these variables. The initial step in optimization involved defining the appropriate space. Utilizing GridSearchCV for model training This contribution involved the systematic execution of the model optimization process.

- I employed GridSearchCV utilizing 5-fold cross-validation for systematic tuning. This procedure systematically trains all potential combinations of parameters within the grid. It employs cross-validation to ensure that the performance score (AUC) is consistent and not reliant on a singular random data partition. The optimization process resulted in the identification of the optimal hyperparameter set for each of the three models, ensuring that the subsequent comparison in Section 5 evaluated the best iteration of each algorithm. This optimization was crucial for the subsequent finding that Logistic Regression demonstrated superior statistical performance.

3. **Ramanaji Boggarapu (11718646):**

Data Preparation and Feature Engineering Role: Developed the foundational data pipeline, ensuring the data was clean, prepared, and appropriately structured for the data mining and modeling phases.

Particular Contributions:

- Initial Cleaning: Conducted the loading, cleaning, and organization of the data for the first instance. This involved addressing missing values in TotalCharges and converting the target variable to a numerical format.

- Preprocessing Pipeline: Employed ColumnTransformer to enhance the robustness and reproducibility of the preprocessing pipeline. This aspect ensured consistency in feature scaling and one-hot encoding methods for subsequent modeling assignments.

4. **Sai Kumar Ramisetti (11822974):**

Conclusive Model Assessment and Statistical Verification Responsibilities included overseeing the final assessment, rigorous validation, and analysis of the trained models.

Particular Contributions:

- Model Evaluation: We executed the optimally tuned models on the reserved test set to obtain the final performance metrics, which comprised the Classification Reports, Confusion Matrices, and AUC values.

- Statistical Validation: We employed the paired t-test on the cross-validation scores of the models to statistically demonstrate that the top-performing model significantly outperformed the rest.
- Model Interpretation: Developed and analyzed interpretative visualizations, including the Decision Tree rules plot and the Logistic Regression coefficient charts, to elucidate the reasoning underlying the models' predictions.