



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

Data Mining and Visualization

Student Name: Suhaana Kapali

Student ID: 2538421

MSc Data and Artificial Intelligence

Module: L7 Data Mining and Visualization

Module Code: ACCA7015

Lecturer: Nough Elmitwally

Table of Contents

1. Introduction.....	4
2. Literature Review.....	4
2.1 Predictive Analysis and Churn Classification	4
2.2 Evaluation Metrics	5
2.3 Clustering for Customer Segmentation	6
2.4 Positioning of this study	6
3. Dataset Description.....	7
4. Methodology	7
4.1 Feature Selection and Preprocessing	7
4.2 Train-test design and scaling.....	8
4.3 Classification Models	9
4.4 Feature Interpretation.....	9
4.5 Clustering Method.....	9
5. Model Implementation and Evaluation	10
5.1 Classification Results	10
5.2 Interpretation of Important Predictors.....	12
5.3 Clustering Results	13
6. Results Analysis and Discussion.....	15
6.1 Comparative Model Performance	15
6.2 Strengths and Limitations of the Modeling Approach	16
6.3 Critical Interpretation of Clustering.....	17
6.4 Integration of prediction and segmentation.....	18
7. Business Recommendations	18
8. Conclusion	20
References	21

Abstract

Customer churn is such a big commercial problem for telecom providers because they lose both the revenue from the customers when they cancel their services and also the cost that would result from replacing a customer who has left. Supervised and unsupervised data mining techniques have been employed here on the Telecom Customer Churn dataset of 100,000 customer records and 100 original attributes. Only 20 numerical and 7 categorical variables were chosen. Missing numbers were replaced with the median, the modes replaced the missing categorical values, and the categorical predictors were transformed via one-hot encoding, producing 47 model inputs. Logistic regression, decision tree, and random forest classifiers were set up and trained through a stratified 80/20 split. The one that achieved the highest test performance when it comes to Random Forest is accuracy 61.68%, precision 60.12%, recall 67.31%, and F1 score 63.52%. Grid search optimization came up with a shorter forest. Yet, its F1-score of 63.44% did not outperform the first model. Feature ranking revealed that the most significant factors contributing to prediction were equipment age, tenure of customer, the average minutes, and other usage changes in the customer's record during recent months. K-Means clustering was used. Then, 49,562 churned customers were grouped per six standard behavioral features. Even though $k = 3$ was chosen from the elbow diagram, the solution resulted in two reasonably large groups and a third one composed of a customer with extreme observations, a cluster with a silhouette score of 0.2594. That's why the cluster analysis results cannot be directly interpreted as three equally representative market segments. Together, the results point to targeted churn scoring, equipment upgrade interventions, and early tenure engagement. But they also point to the need for more robust outlier treatment and deeper service quality data before practical implementation.

1. Introduction

Churn is the term used to describe when a subscriber cancels or does not renew a contract with a service provider [1]. Churn is particularly relevant in telecommunications, as customers may typically compare pricing and change providers and transfer numbers with relative ease. A provider that can identify likely churners before cancellation can direct retention resources toward customers who are most at risk and most valued. Predictive analytics, then, translates past customer behavior into forward looking data for marketing, customer service, and revenue protection decisions.

The assessment includes the implementation of at least two classification approaches and either clustering or association rule mining on the same telecom dataset used in the preceding exploratory assignment. It also asks for critical model comparisons, analysis of findings, and actionable recommendations. The supplied notebook fulfills this need by developing three classifiers, logistic regression, decision tree, and random forest, followed by K-means clustering of customers whose churn value is 1.

The purpose is to find out how well revenue, usage, account, device, and demographic data may predict customer retention status and to investigate if churned customers constitute meaningful behavioral groupings. The objectives are to build a modelling dataset, to train and evaluate classification algorithms, to optimize the best collective model; to analyze coefficients and feature importance; to cluster the churning population; and to translate findings into actionable retention suggestions.

2. Literature Review

2.1 Predictive Analysis and Churn Classification

Predictive analytics employs past data and statistical or machine learning algorithms to guess at an unknown or future outcome [1]. Churn prediction is usually posed as a binary classification problem, where kept customers are classified as 0 and churned customers as 1. Logistic regression is still a valuable baseline method, as the coefficients are directly interpretable as to whether a predictor improves or decreases the predicted log-odds of churn [2]. It is simple, which encourages transparency, but the model assumes a linear relationship between the predictors and the log-odds of the goal. Complex customer

behaviour may break this assumption, as turnover may be caused by interactions between tenure, changing consumption, device age and costs.

Decision Trees recursively partition the feature space into more homogeneous sections by identifying split points that minimize the class impurity [3]. The key advantage is interpretability. A path through the tree can be represented as a collection of decision rules. But a single tree can be unstable and overfit the training data, especially if there are numerous plausible splits. Therefore, maximum depth and minimum split size should be limited. Random forest solves instability by training numerous trees using bootstrap samples and random selections of predictors and then averaging their predictions [4]. This often decreases variance and increases generalization on organized tabular data. In recent telecom studies, it is widely reported that ensemble trees outperform individual linear or tree models, with the degree of improvement highly dependent on data quality and feature relevance.

Random Forest models have better predictive performance than individual classifiers, but they are also less interpretable, as the final prediction is based on the collective judgment of many trees instead of a single set of rules. This results in a trade-off between forecast accuracy and model transparency in business domains like telecommunications. Organizations may prefer models with slightly lower accuracy but better interpretability when client retention choices need to be justified. So we have a balanced assessment of the prediction capabilities and the real business applicability by comparing Logistic Regression, Decision Tree and random forests.

2.2 Evaluation Metrics

Accuracy is the percentage of all test observations that were correctly classified [5]. This is instructive because the churn is nearly evenly split: 50.44% of customers stayed and 49.56% churned. Precision is the ratio of projected churners that churned, and recall is the ratio of actual churners that were spotted properly. The F1-score is the harmonic means of precision and recall and is effective when both missed churners and superfluous retention offers are important. Recall can also be emphasized in a commercial situation, because a false negative means that a true churner is not receiving an intervention. However, very low precision would waste retention resources; hence, model selection

should be based on the precision–recall tradeoff, not only on accuracy. In addition to these metrics, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) provide threshold independent measures of classification performance. Higher AUC values indicate better discrimination between churned and retained customers, making ROC-AUC a useful complement to accuracy and F1-score when comparing classification models.

2.3 Clustering for Customer Segmentation

Classification answers the question of “who is likely to churn?” and clustering analyzes whether customers have distinct patterns without a known target label. K-Means is popular due to its computational efficiency and the translation of its centroids into segment profiles [6]. It minimizes within cluster squared Euclidean distance which requires standardization when variables are measured in various units. Its restrictions are also relevant. K must be pre-specified, outcomes can be sensitive to starting centroids and extreme observations can drag centroids away from the majority. The elbow method and silhouette coefficient are common evaluation techniques, but neither ensures that the clusters have business meaning [6]. A silhouette score near to 1 suggests good separation, about 0 implies overlap and negative values possibly improper portion. Moderate scores are prevalent in real customer data. However, even small clusters should still lead to analysis and not immediate business branding.

2.4 Positioning of this study

The study is based on a two stage design as per the principles of CRISP-DM: data understanding and preparation, classification, evaluation, and customer segmentation [7]. The design is suitable in that it combines an operational prediction task with an exploratory explanation task. Crucially, this research does not assume the validity of all grouping outputs. The notebook produced an unusual one customer cluster, which is closely analyzed as evidence of an extreme observation dominating the three cluster solution.

3. Dataset Description

The Telecom Customer Churn data collection has 100,000 rows and 100 columns [8]. The aim variable is churn with 50,438 maintained customers and 49,562 churned consumers. No duplicate rows were found. Original variables are monthly revenue, minutes of use, overage activity, dropped or blocked calls, customer care interactions, account tenure, total calls and revenue, device price and age, household characteristics, and demographic categories.

The missingness is focused on the demographic factors. The notebook shows 49,366 missing values for number of automobiles, 38,308 for dwelling size, 37,923 for household status, 33,706 for ownership/rental status, and 31,909 for dwelling type. Most of the fundamental usage and income factors have far fewer missing data. This distinction is significant since transactional aspects are automatically captured, whereas demographic data could be optional or incomplete. The modeling strategy preserves some of the demographic data but imputes them, thereby retaining the sample size but at the cost of reducing the ambiguity as to whether the fact of missingness itself is informative [9].

Class	Customers	Percentage
Retained (0)	50,438	50.44%
Churned (1)	49,562	49.56%

4. Methodology

4.1 Feature Selection and Preprocessing

Twenty numerical predictors were chosen: rev_Mean, mou_Mean, totmrc_Mean, da_Mean, ovrmou_Mean, ovrrrev_Mean, roam_Mean, change_mou, change_rev, drop_vce_Mean, blk_vce_Mean, custcare_Mean, threeway_Mean, months, totcalls, totmou, totrev, hnd_price, phones, eqpdays. Also kept are seven categorical predictors: marriage, ethnic, creditcd, new_cell, refurb_new, hnd_webcap, and ownrent. These factors are revenue, service use, behavioral change, call quality, account length, device attributes, and chosen demographics. Customer_ID was removed, as it is an identifier and not a behavioral cause of churn.

Numerical missing values were converted to numeric data types when necessary and replaced by the median. Median attribution is strong to extreme values and was therefore appropriate for biased telecommunication variables. Categorical missing values were substituted with mode and converted to string values. Categorical variables were converted into binary indicators using one hot encoding, dropping the first category [10]. The resulting predictor matrix included 47 columns and had no missing data. This approach is simple and reproducible, but there is a minor danger of information leaking when attribution is performed prior to the train-test split, as medians and modes were computed on the entire dataset [11]. In a stricter production approach, complaint and encoding would be placed inside a pipeline trained on the training data exclusively.

Preprocessing data is one of the most crucial stages in any machine learning project, as poor quality of input data typically results in inconsistent model performance. We chose median attribution over mean imputation because some numerical variables, such as customer revenue and call time, have a wide range and may contain outliers. The median values are less sensitive to outliers and so retain a more representative central trend. Likewise, one-hot encoding represents categorical variables as binary indicators, avoiding the false assumption of a numerical relationship between categories. The data can then be used in classification methods.

4.2 Train-test design and scaling

The data were split into 80,000 training rows and 20,000 testing rows using `random_state = 42` and stratification on churn. Stratification preserved the near-balanced target distribution in both subsets. Standard Scaler was fitted to the training predictors and applied to both training and test data for logistic regression. Decision Tree and random forests were trained on unscaled features because tree splits depend on ordering and thresholds rather than Euclidean distance or coefficient magnitude [12].

To increase the reliability of the model performance evaluation, it is useful to use a stratified train-test split, which preserves the same proportion of churned and retained consumers in both the training and testing datasets [12]. The only algorithm where feature scaling was performed is Logistic Regression, since this technique is sensitive to the changes in the scale of the predictor variables. On the other hand, Decision Tree and

Random Forest algorithms are not influenced by feature scaling as they do not rely on distance calculations but instead use decision thresholds to split the data [14].

4.3 Classification Models

Logistic Regression was set with `max_iter=1000`, `random_state=42`. The Decision Tree uses `max_depth=10` and `min_samples_split=20` to limit overfitting. The first Random Forest was built with 100 trees, `max_depth = 12`, `min_samples_split = 10` with parallel processing. All models were evaluated on the identical held out test set using accuracy, precision, recall, F1-score, confusion matrices, classification reports, and ROC-AUC analysis. These complementary metrics provide a more comprehensive assessment of classification performance by evaluating both predictive accuracy and the models' ability to distinguish between churned and retained customers.

Then, we applied a GridSearchCV technique to Random Forest using three-fold cross-validation and F1 as the scoring criterion. The candidate grid varied `n_estimators`, `max_depth` and `min_samples_split` on the candidate grid: The optimal cross-validation setup was 100 trees, `max_depth = 10` and `min_samples_split = 20` with a mean cross-validation F1-score of 0.6348. Once tweaked, the model was tested on the unmodified test set.

4.4 Feature Interpretation

The Logistic Regression coefficients were ranked by absolute magnitude to identify variables most strongly associated with the model decision. Random Forest feature importance was also extracted. Coefficients provide direction but should not be interpreted causally, especially when predictors are correlated. Forest importance indicates how much a feature contributes to impurity reduction across trees, but it can favour continuous variables or variables with many possible split points. The two views are therefore interpreted as complementary rather than definitive causal evidence [4] [12].

4.5 Clustering Method

Clustering was restricted to churned customers and used `rev_Mean`, `mou_Mean`, `change_mou`, `change_rev`, `months` and `eqpdays`. Missing numerical values were median-imputed and all six variables standardised. K-Means models were fitted for `k = 1` to `10`, and `k = 3` was selected visually from the elbow plot. The final model used `random_state`

= 42 and $n_{\text{init}} = 10$. Cluster validity was assessed with the silhouette score, and Principal Component Analysis reduced the six dimensions to two for visualisation. PCA was not used to train K-Means; it served only as a display technique.

5. Model Implementation and Evaluation

5.1 Classification Results

Model	Accuracy	Precision	Recall	F1-Score
logistic regression	57.94%	57.50%	57.98%	57.74%
Decision Tree	59.95%	59.37%	60.77%	60.06%
Random Forest	61.68%	60.12%	67.31%	63.52%
Tuned Random Forest	61.21%	59.52%	67.91%	63.44%

Random Forest gave the best untuned accuracy and strongest F1-score. It was able to recall 67.31% of the real churners, which is nearly two-thirds, whereas the Decision Tree and Logistic Regression were able to recall 60.77% and 57.98% respectively. The increase in performance from Logistic Regression to tree-based models demonstrates that non-linear interactions are important. However, the accuracy of 61.68% is still poor, suggesting that the factors chosen to do not properly discriminate between churners and retained consumers.

The modified Random Forest somewhat improved recall to 67.91% but lowered precision and accuracy, resulting in its F1-score being slightly lower than the original forest. This is an important result: cross-validation tuning did not ensure a better held out result. The difference is slight but favors keeping the original Random Forest as the recommended model for this test set. To argue that a single configuration generalizes reliably, a broader tiered validation architecture would be required [15].

It should be noted that none of the categorization models obtained very high prediction accuracy despite very large dataset used. This implies that there are many behavioral and external elements that drive customer churn that are not fully captured in the provided dataset. Factors like customer complaints, service quality, network performance, competition promotions and customer happiness are not covered but will probably be important in churn behavior. Therefore, enhancing the data quality might have a larger impact on the prediction performance than merely using more complicated algorithms [16]. ROC curve analysis provided additional evidence supporting the model comparison. As shown in Appendix H, the Random Forest achieved the highest AUC score (0.667), followed by the Decision Tree (0.640) and Logistic Regression (0.610). Although all three models demonstrated only moderate discrimination, the Random Forest consistently produced the strongest overall classification performance across both threshold dependent and threshold independent evaluation measures.

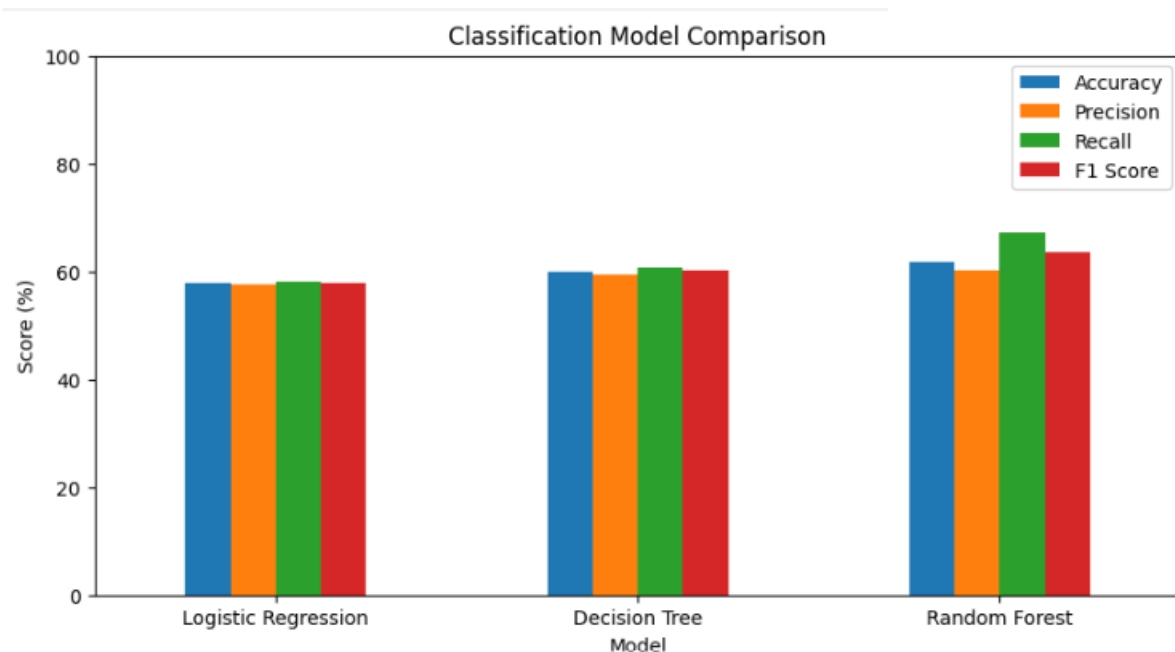


Figure 1 CLASSIFICATION MODEL COMPARISON

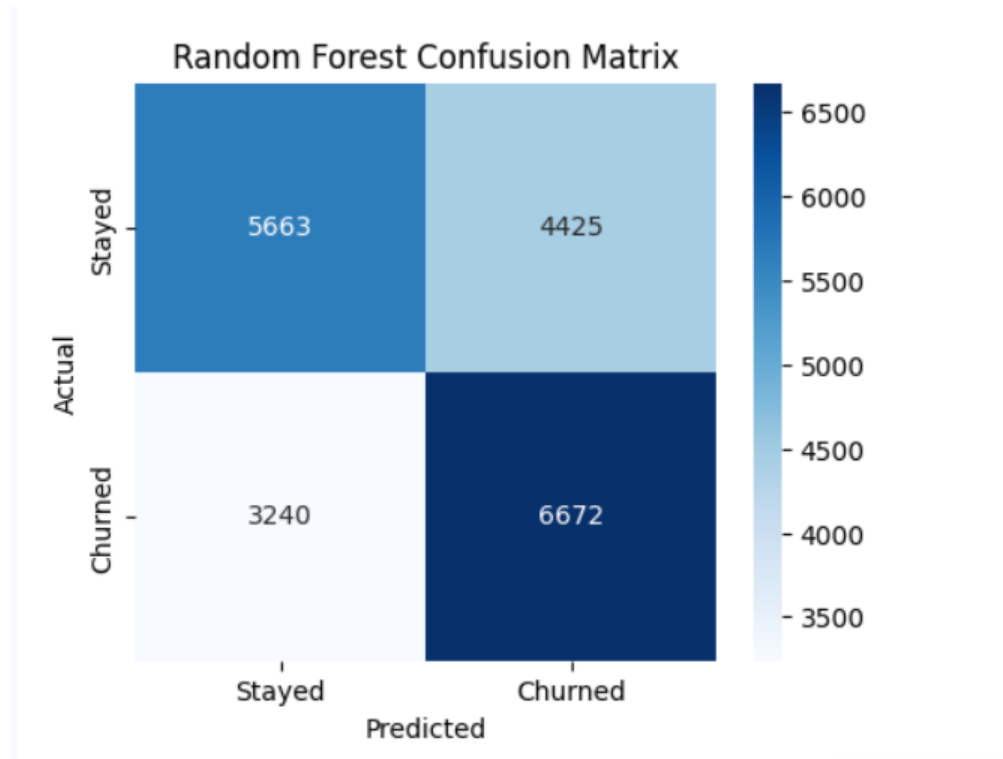


Figure 2 RANDOM FOREST CONFUSION MATRIX

5.2 Interpretation of Important Predictors

The largest absolute coefficients for Logistic Regression were total minutes, average monthly minutes, equipment age, account months, selected ethnicity markers, monthly revenue, recent revenue change and recent use change. Positive coefficients raise simulated log-odds of turnover after accounting for other variables; negative coefficients decrease them. Total minutes had a positive coefficient, and average monthly minutes had a negative coefficient, showing the need for consideration of correlated measurements and varied time frames for coefficient interpretation [17].

Random Forest found equipment age to be the most relevant predictor (0.1758), followed by account tenure (0.1476), average monthly minutes (0.0720), change in minutes (0.0639), total monthly recurring charge (0.0593), handset price (0.0490), and total revenue (0.0486). The results of this ranking indicate that customer relationship and device lifetime timing have a crucial role in discriminating churn. An aging device can make competitor upgrade offers more convincing, whereas tenure can capture stages of

the contract, loyalty, or cumulative unhappiness. Recent change variables are relevant especially since falling engagement may manifest itself before cancellation.

The rankings of relevance also suggest that changes in consumer behaviour are often better predictors of churn than static demographic information. Variables such as change in monthly consumption and change in revenue also provide evidence of recent behavioural changes, demonstrating that customers often show warning indications before they quit their account. By real-time monitoring of these behavioral variables, telecommunication providers could intervene early with tailored marketing campaigns or personalized retention offers.

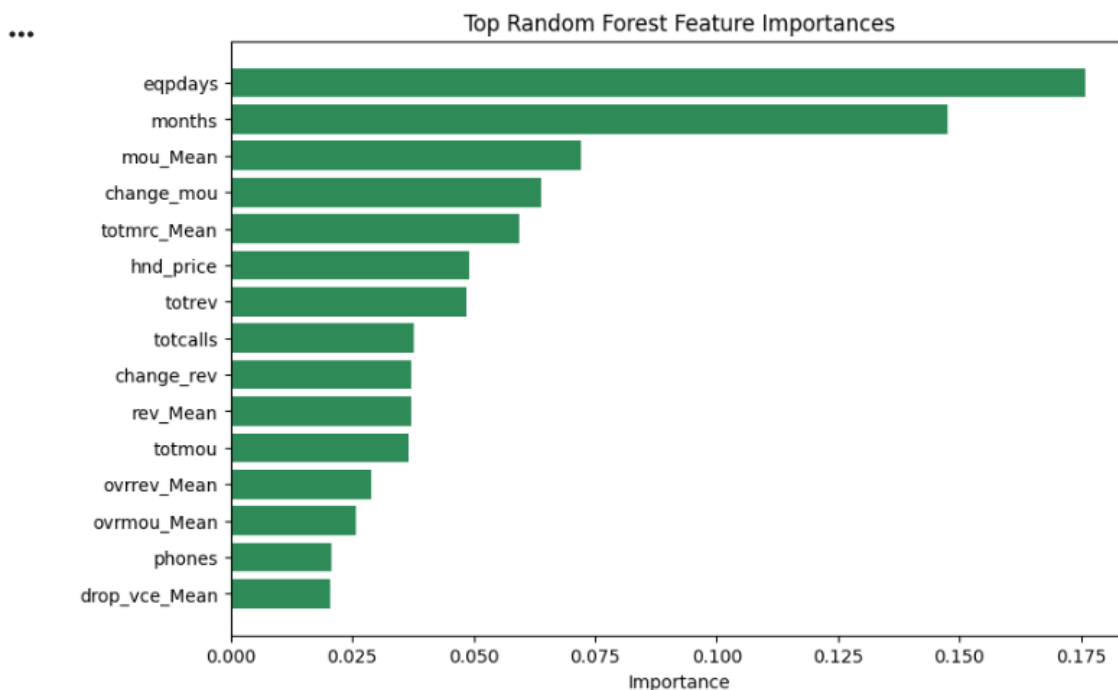


Figure 3 TOP RANDOM FOREST FEATURE IMPORTANCES

5.3 Clustering Results

The churned subset contained 49,562 customers. The overall averages were £58.16 for monthly revenue, 482.55 monthly minutes, a change in minutes of -22.58, a change in revenue of -0.62, tenure of 19.04 months and equipment age of 421.09 days. The elbow plot suggested three clusters, and the resulting silhouette score was 0.2594, indicating overlapping and only weak to moderate structure.

Cluster	Assigned label	Customers	rev_Mean	mou_Mean	change_mou	change_rev	months	eqp-days
0	Long Tenure Churners	16,806	43.94	255.42	-16.30	-0.60	28.16	681.21
1	Low Usage Churners	32,755	65.35	598.85	-27.00	-0.94	14.36	287.64
2	High Value Churners	1	3,843.26	12,206.75	31,219.25	9,963.66	16.00	54.00

From the profile table a severe validity concern is revealed. Cluster 2 is merely one consumer with extraordinary values well above the rest of the population. This is not a steady consumer group; it is an isolated multivariate outlier, captured as its own centroid. As a result, naming them High Value Churners could mislead decision makers to believe that the high-value category is sizable. The meaningful structure is mostly between Cluster 0, with longer tenure and older devices, and Cluster 1, with the majority, shorter tenure, newer equipment, and higher utilization. Even the designation of Low Usage Churners is relative and imperfect as Cluster 1 has a higher average usage compared to Cluster 0. Labels should consequently be updated following strong outlier management and re-clustering.

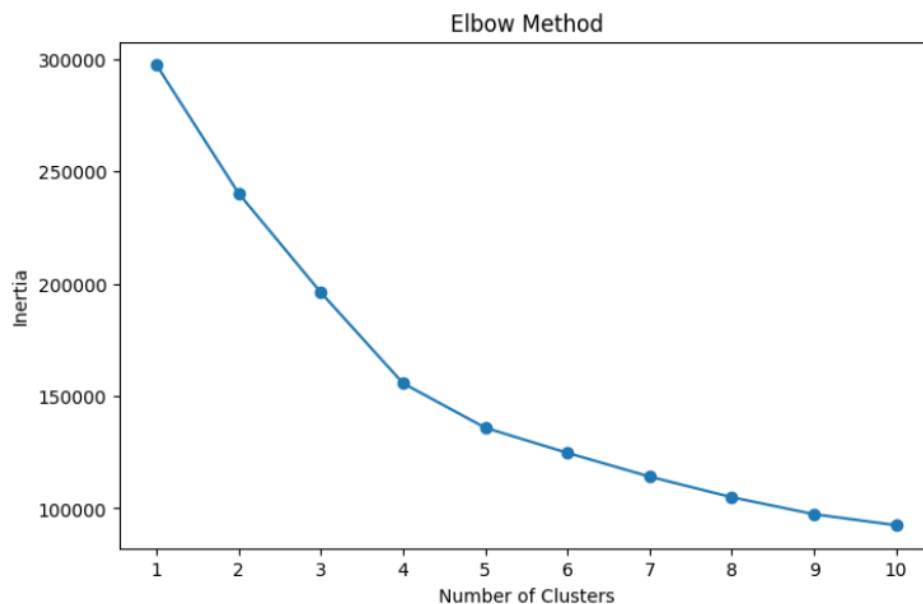


Figure 4 ELBOW METHOD FOR K=1-10W

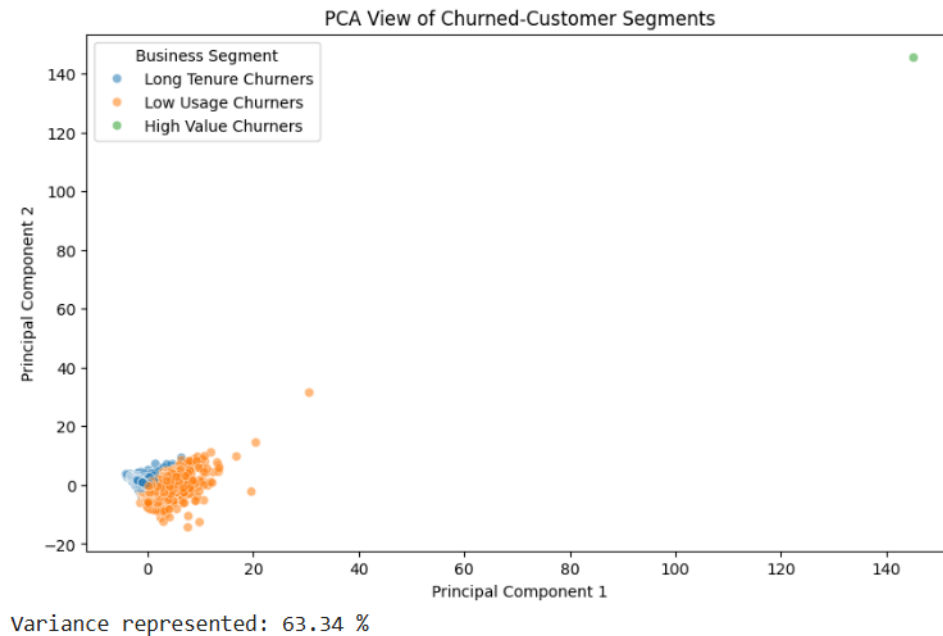


Figure 5 PCA visualisation of churned-customer clusters; the first two components explain 63.34% of variance

6. Results Analysis and Discussion

6.1 Comparative Model Performance

The order of model performance is theoretically plausible. Logistic regression produced a linear border and had the worst F1-score. Decision trees were able to model non-linear thresholds and enhanced accuracy and F1 score [18]. Random Forest greatly enhanced its performance by averaging several trees minimizing the instability of a single tree. The 3.46 percent F1 improvement from Logistic Regression to Random Forest is notable but not large enough to demonstrate that algorithm choice alone solves churn prediction. The low ceiling is consistent with studies that have shown that transactional data typically misses crucial factors such as competitor offers, network issues, complaint content, contract events and customer mood.

The Random Forest confusion matrix demonstrates a purposeful or promising bias to discover churners, with memory for class 1 being higher than memory for class 0 [19]. This may be desired when the cost of a missed churner is greater than the expense of

addressing a retained client. But the business should work out real campaign economics. For example, if an offer is pricey, precision might need to improve; if a customer who churned has a high lifetime value, recall might be more important. Using the anticipated probability rather than the accepted 0.5 threshold to tune the threshold might get a better commercial trade off.

From a commercial viewpoint, the selection of the highest performing model should not be based on statistical performance alone. In addition, organizations must evaluate the cost of implementation, computing requirements, model maintenance, and simplicity of interpretation [20]. Logistic Regression is computationally efficient and easy to explain to corporate management. In contrast, Random Forest uses larger computer resources, but typically gives superior prediction performance. The ultimate model selection is, thus, a trade-off between operational efficiency and predictive performance [21].

6.2 Strengths and Limitations of the Modeling Approach

The analysis offers several strengths. It employs a huge dataset, a stratified hold-out test set, numerous algorithms, four evaluation metrics, cross-validation tuning, and feature-level interpretation. It also fulfills the condition of a combination of classification and clustering for the assessment. The same test set allows for a fair comparison, and the balanced objective prevents the misleadingly high accuracy that can be achieved with severely biased churn data.

Several constraints affect the strength of the conclusions. Preprocessing statistics were first computed prior to splitting, producing possible leakage. Secondly, one-hot encoding was performed outside a machine learning pipeline, making it more difficult to implement on unseen data. Third, just a limited hyperparameter search space was investigated. ROC curve analysis and AUC evaluation were conducted to provide further insight into model discrimination (see Appendix H). Probability measurement and precision-recall curve analysis were not conducted. Furthermore, feature importance is not causal and may be biased toward variables with multiple potential split points. Finally, numerous demographic variables had large missing values and were imputed with the mode, which may lessen variability within those features. Future study should also assess fairness

across demographic groups and consider eliminating sensitive variables that have low predictive value.

Another limitation is the evaluation technique. Here, model performance is evaluated using a single train-test split. This is a conventional strategy, but it may not reflect the variation of model performance across different samples. Other techniques, e.g., repeated k-fold cross-validation or nested cross-validation, may yield more reliable estimates of model generalization. Moreover, future studies could explore ensemble stacking, gradient boosting approaches e.g. XGBoost, LightGBM and probability measurement for additional improvement in predicting performance.

6.3 Critical Interpretation of Clustering

The clustering step explains why unsupervised findings require more rigorous validation than a scatter plot that looks acceptable. The elbow approach recommended $k=3$, yet one cluster has only one consumer. K-Means minimizes squared distance so one extreme point can dominate the objective and form an isolated cluster. The silhouette score of 0.2594 also shows a large amount of overlap [22]. Hence, the answer should be seen as exploratory evidence and not as a definitive segmentation scheme. Therefore, the findings of the clustering should be handled with caution. The elbow technique revealed three clusters, however the presence of an isolated single customer cluster indicates that the optimization criteria are strongly influenced by an extreme observation. It can be observed that one of the major drawbacks of the K-Means technique, which is sensitivity to outliers [22]. Before using clustering in a genuine business scenario, you might wish to explore other pre-processing methods, such as outlier removal or robust scaling, to get more stable and meaningful customer clusters.

A more robust procedure would be to check and minorize or reliably scale extreme values, compare $k = 2$ and $k = 3$, assess cluster stability across random seeds, and test other algorithms such as K-Medoids, DBSCAN, or Gaussian Mixture Models. The present output still indicates that the two main groups are useful hypotheses: that one group is older with longer tenure and older equipment, while the other is newer and more active. The patterns can inform pilot actions, but the one-customer cluster should not support a broad campaign.

6.4 Integration of prediction and segmentation

Classification and clustering have separate but complementary goals. Random Forest can rank active customers by churn risk, and validated customer profiles can dictate the sort of retention action. An operational workflow in the future could score clients monthly, classifying them into a robust behavioral profile and selecting offers based on the predicted value. However, the current clustering approach still lacks the reliability needed for automated assignments. The initial utilization would thus be a risk score coupled with simple, interpretable business rules based on equipment age, tenure, and falling usage.

7. Business Recommendations

- Deploy the original Random Forest as the current benchmark model, not the end system without inquiry. It achieved the highest held-out F1-score of 63.52% and a high recall of 67.31%. Before you start a complete campaign, predictions should be tested in a trial campaign.
- Make use of probability, rather than a set binary decision. The provider can then apply scarce retention resources to high-risk, high-value accounts based on predicted churn likelihood and expected lifetime value.
- Design an intervention for equipment aging. Since eqpdays was the strongest Random Forest predictor and the longer-tenure cluster had an average of 681 equipment days, consumers near a set device-age threshold should be presented with upgrade, trade-in, or loyalty offers before contract renewal or switching.
- Get customers involved early in the relationship 12 to 18 months. The largest cluster lasted on average only 14.36 months. Early service recovery contacts,

usage instruction, plan fit checks, and welcome calls can help reduce the disengagement before loyalty takes hold.

- Start outreach for unexpected use spikes / revenue surges. Change_mou and change_rev were included in both our coefficient analysis and importance analysis. Sudden declines should signal diagnostic communication; unexpected charges should trigger bill explanation or plan right-sizing support.
- After treating the outliers, repeat the clustering step; The one-customer cluster should be investigated as a probable data error, atypical corporate account, or a real extreme user. It needs to be dealt with on an individual basis before you can commercially name or exploit client segments.
- Improve future data collection. Billing and use characteristics may explain churn less directly than complaint grounds, network quality, dropped call location, contract expiry, competitive offers, customer-service text and promotion history. These characteristics can both enhance the forecasting performance and make the recommendations more actionable.
- Establish governance and fairness controls. Monitor model performance for drift, tuning and subgroup error rates across time. If it is shown that sensitive demographic predictors contribute to injustice with no relevant performance advantage, justify, record and remove them.

8. Conclusion

In this paper, the utilization of logistic regression, decision tree, and random forest models on a balanced dataset for telecommunications churn and added K-means clustering of churned customers to supplement the classification task was done. The Random Forest classifier performed the best, with an accuracy of 61.68%, precision of 60.12%, recall of 67.31%, and F1-score of 63.52%. A modest improvement was made in recall, but the overall F1 score on the test set was not improved. We found that the most important factors were equipment age, tenure, average monthly usage, and recent behavioral changes.

The clustering study has a silhouette score of 0.2594, and two huge groups of customers were obtained; however, the third cluster had only one extreme consumer. This outcome undermines the three segment interpretation proposed and emphasizes the necessity of a critical assessment of unsupervised models rather than taking labels at face value. The research provides a solid baseline for churn-risk prioritization and highlights practical intervention themes around device updates, early tenure assistance, and behavioral change. The pipeline needs to be rebuilt as a leakage safe pipeline before usage in production, evaluated with probability based business KPIs, tested for fairness, and supported with better service quality data.

The research generally indicates the potential of data mining technologies to support the evidence based decision making in the telecoms business. The predictive performance is moderate, but the study manages to identify key behavioral characteristics associated with customer churn and shows how supervised and unsupervised learning can be combined to generate significant business insights. Future work should be incorporated. More data regarding customer interactions, improved feature engineering, more advanced collection approaches, and validation of the models on different data sets. The improvements can lead to more precise prediction systems and more dependable client segmentation strategies for long term retention planning.

References

- [1] V. H. K. Q. A. A. F. O. Chang, "Prediction of Customer Churn Behavior in the Telecommunication Industry Using Machine Learning Models," *Algorithms*, vol. 17, no. 6, p. 231, 2024.
- [2] D. W. L. S. & S. R. X. Hosmer, *Applied Logistic Regression (3rd Edition)*, John Wiley & Sons, 2013.
- [3] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning*, vol. 1, no. 1, p. 81–106, 1986.
- [4] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, p. 5–32, 2001.
- [5] s.-l. Developers, "Model Evaluation: Quantifying the Quality of Predictions," scikit-learn, 2025. [Online]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html. [Accessed 20 07 2026].
- [6] A. M. A. A. O. E. A. E. e. a. Ikotun, "K-Means Clustering Algorithms: A Comprehensive Review," *Algorithms*, 2023.
- [7] P. J. J. K. T. Chapman, *CRISP-DM 1.0 Step-by-Step Data Mining Guide*, SPSS, 2000.
- [8] L. & S. Hosmer, *Applied Logistic Regression (3rd ed.)*, N/D, 2013.
- [9] R. J. A. R. D. B. Little, *Statistical Analysis with Missing Data*, John Wiley & Sons, 2019.
- [10] R. J. A. Little, *Statistical Analysis with Missing Data*, John Wiley & Sons, 2019.
- [11] F. V. G. Pedregosa, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, p. 2825–2830, 2011.
- [12] G. W. D. H. T. & T. R. James, *An Introduction to Statistical Learning: With Applications in Python (2nd ed.)*, Springer, 2021.
- [13] F. V. G. G. A. M. V. T. B. G. O. B. M. P. P. W. R. D. V. V. J. P. A. C. D. B. M. P. M. & D. É. Pedregosa, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, p. 2825–2830, 2011.
- [14] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, p. 5–32, 2001.

- [15] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 1995.
- [16] P. Domingos, "A Few Useful Things to Know About Machine Learning," *Communications of the ACM*, vol. 55, no. 10, p. 78–87, 2012.
- [17] R. X. Sturdivant, *Applied Logistic Regression*, John Wiley & Sons, 2013.
- [18] W. Verbeke, "New Insights into Churn Prediction in the Telecommunication Sector: A Profit Driven Data Mining Approach," *European Journal of Operational Research*, vol. 218, no. 1, p. 211–229, 2012.
- [19] M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLOS ONE*, vol. 10, no. 3, 2015.
- [20] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*, vol. 1, no. 5, p. 206–215, 2019.
- [21] C. Molnar, *Interpretable Machine Learning* (2nd ed.), Lulu.com, 2022.
- [22] P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computational and Applied Mathematics*, vol. 20, p. 53–65, 1987.
- [23] A. Jain, "Data Clustering: 50 Years Beyond K-Means," *Pattern Recognition Letters*, vol. 31, no. 8, p. 651–666, 2010.

Appendix

Appendix-A Selected Features

Numerical Features

- rev_Mean
- mou_Mean
- totmrc_Mean
- da_Mean
- ovrrou_Mean
- ovrrev_Mean
- roam_Mean
- change_mou
- change_rev
- drop_vce_Mean
- blk_vce_Mean
- custcare_Mean

- threeway_Mean
- months
- totcalls
- totmou
- totrev
- hnd_price
- phones
- eqpdays

Categorical Features

- marriage
- ethnic
- creditcd
- new_cell
- refurb_new

- hnd_webcap
- ownrent

Appendix B- Dataset Summary

Item	Value
Rows	100,000
Columns	100
Training	80,000
Testing	20,000
Predictors	47
Churned	49,562
Retained	50,438

Appendix C- Missing Value Treatment

- Numerical variables: Median imputation
- Categorical variables: Mode imputation
- One-hot encoding applied
- Final dataset contained no missing values

Appendix-D Model Performance

Model	Accuracy	Precision	Recall	F1
Logistic Regression	57.94%	57.50%	57.98%	57.74%
Decision Tree	59.95%	59.37%	60.77%	60.06%
Random Forest	61.68%	60.12%	67.31%	63.52%

Tuned Random Forest	61.21%	59.52%	67.91%	63.44%
------------------------------------	--------	--------	--------	--------

Appendix E – GridSearchCV Results

Best Parameters:

n_estimators = 100

max_depth = 10

min_samples_split = 10

Cross-validation F1-score = 0.635687

Appendix F – Random Forest Feature Importance

Feature	Importance
Equipment age	0.1758
Account tenure	0.1476
Average monthly minutes	0.0720
Change in minutes	0.0639
Monthly recurring charge	0.0593
Handset price	0.0490
Total revenue	0.0486

Appendix G – Clustering

Algorithm: K-Means

Number of clusters: 3

Silhouette score: 0.2594

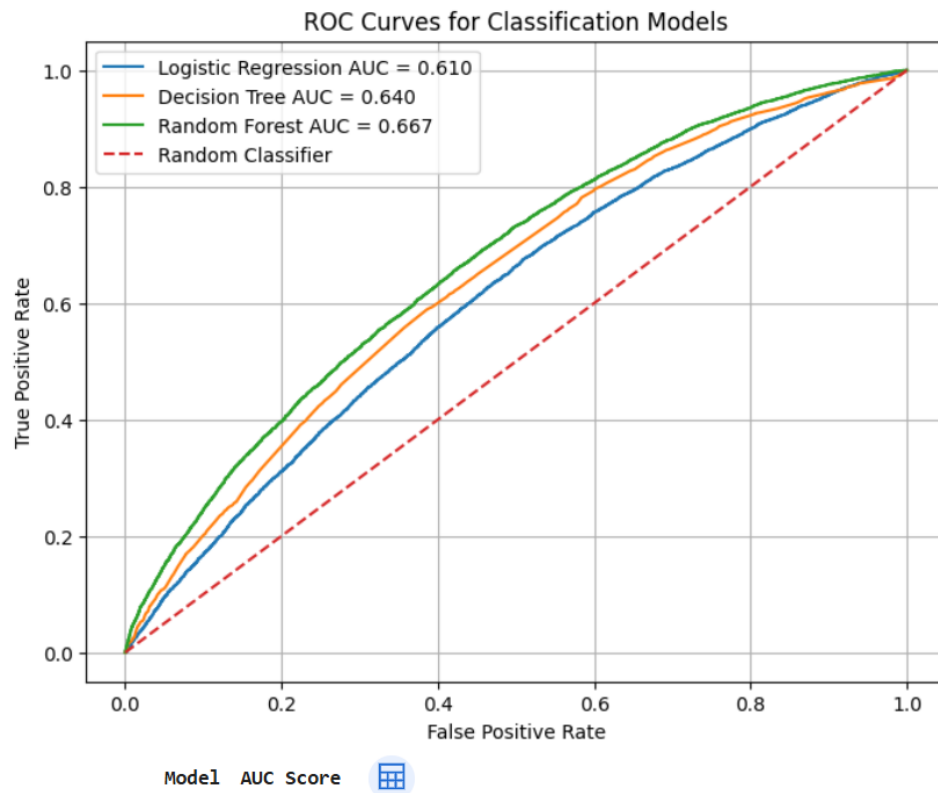
PCA variance explained: 63.34%

Cluster sizes:

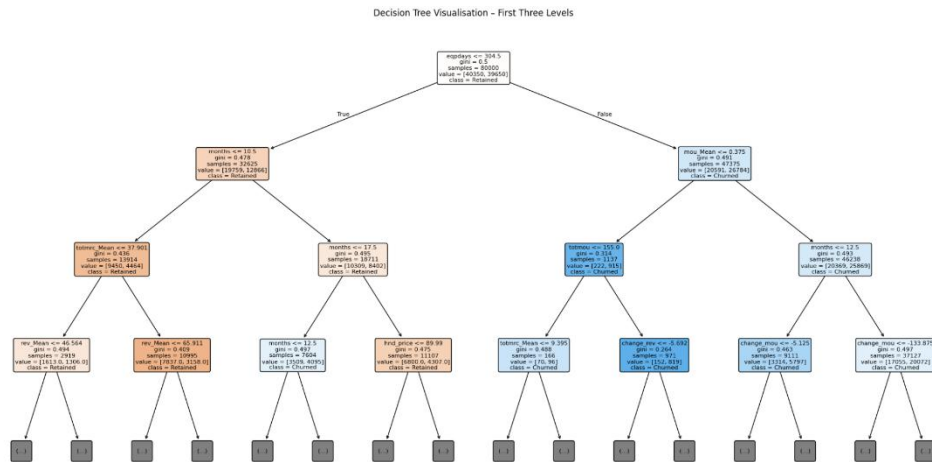
Cluster 0: 16,806

Cluster 1: 32,755

Cluster 2: 1

Appendix H – ROC Curves

Appendix I – Decision Tree Visualization



Appendix J- Hyperparameter Tuning Results

	Number of Trees	Maximum Depth	Minimum Samples Split	Mean CV F1 Score	Standard Deviation	Rank
0	100	10	10	0.635687	0.001700	1
1	150	10	10	0.635274	0.001028	2
2	150	10	20	0.633801	0.000477	3
3	100	10	20	0.633517	0.000346	4
4	150	15	20	0.632892	0.001820	5
5	150	15	10	0.632096	0.001314	6
6	100	15	20	0.630769	0.002216	7
7	100	15	10	0.630758	0.000829	8

Appendix K - Software

- Python
- Google Colab
- Pandas
- NumPy
- Scikit-learn
- Matplotlib
- Seaborn