SWINBURNE UNIVERSITY OF TECHNOLOGY

MASTER OF DATA SCIENCE

COS60008 - Introduction to Data Science

2024, Semester 1

Customer Churn Analysis: Telecom data set

Name: Thi Ngan Ha Do

Student ID: 103128918

Email Address: 103128918@student.swin.edu.au

Submission Date: 11/05/2024

1. Abstract

This report delves into churn analysis within the telecom business domain, focusing on predicting and building churn customers profiles to identify common traits in the dataset.

To achieve this, the report employs data preparation, visualization, and classification modelling techniques, primarily utilizing K-Nearest Neighbors (KNN) and Random Forest (RF) algorithms. The process includes data preparation, data exploration, and data preprocessing which includes standardization, feature selection, and hyperparameter tuning to enhance model performance and reduce overfitting.

The key outcomes of the study reveal that Suite 2 demonstrated the highest recall, indicating its effectiveness in capturing actual churn customers. Furthermore, KNN was preferred for its ability to prioritize capturing churn customers over avoiding false positives.

2. Introduction

This report presents the results of Tasks 1, 2, and 3, focusing on segmenting customers at risk of churn and training classification models to improve churn prediction. Through extensive data visualization and modeling, we identified lodged complaints, activity status, phone service usage, and customer value as the primary indicators of churn. Furthermore, our analysis demonstrates that both the KNN and RF classification models exhibit high performance efficiency and versatility for various applications.

3. Task 1, Task 2, Task 3

3.1 Task 1 - Problem Formulation, Data Acquisition and Preparation

3.1.1 Data Acquisition

‘Customer Churn’ CSV file was loaded using pandas' "read_csv" function. Subsequently, ".equals" method was applied to ensure that the loaded dataframes matched the raw data files. The result returned "True", showing that the data was correctly and completely sent from the source data file.

3.1.2 Data Preparation

Error Type	Columns	Issues	Detection	Solution
Data types	Binary: Complains, Tariff Plan, Status, Churn Ordinal: Charge Amount, Age Group	Int data types	Used `df.info()` to check data types of all columns	Converted the data types of binary columns to 'object' for visualization purposes, then reverted them back to 'int' for data modelling. The data types of ordinal columns remained unchanged, as they have a hierarchical structure.
Extra whitespaces	Call Failure; Subscription Length; Charge Amount; Seconds of Use; Frequency of use; Frequency of SMS; Distinct Called Numbers; Age Group; Tariff	Extra whitespaces in column names	Used `df.head()` to get a broad overview of the data.	Applied `str.replace` to remove inconsistent whitespaces, ensuring smooth visualization and easier access to columns in subsequent code.

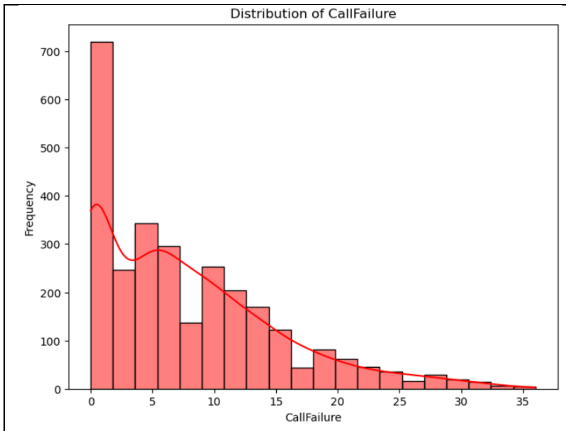
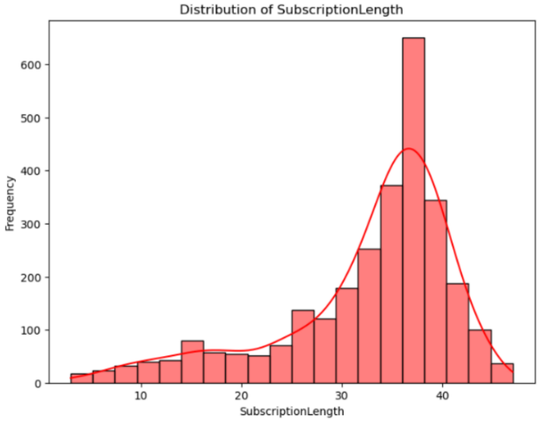
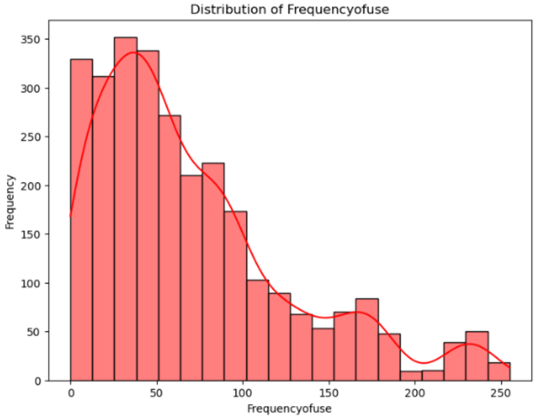
	Plan; Customer Value			
Duplicates	Dataset does not have a unique identifier column	300 duplicates	Used `.duplicated()` function to detect duplicates	Applied `drop_duplicates` considering the low likelihood of natural occurrences (given all columns are in numerical forms), suggesting a potential data entry issue

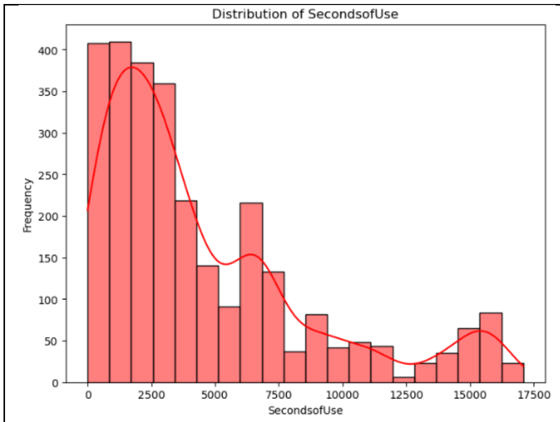
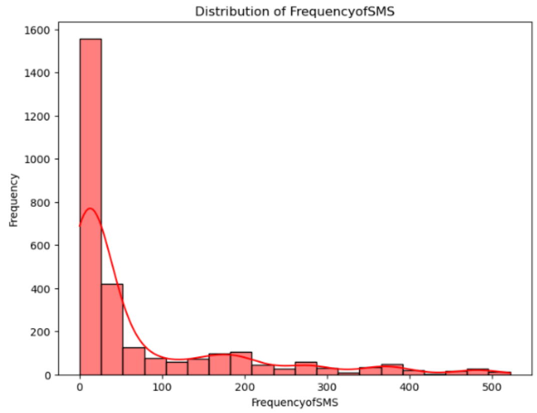
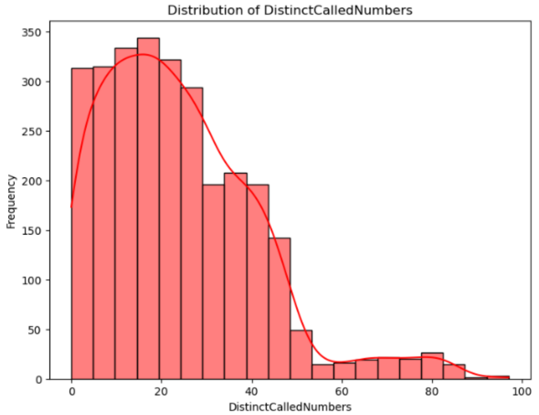
Table 1 summarises the detection and solution for data cleaning issues

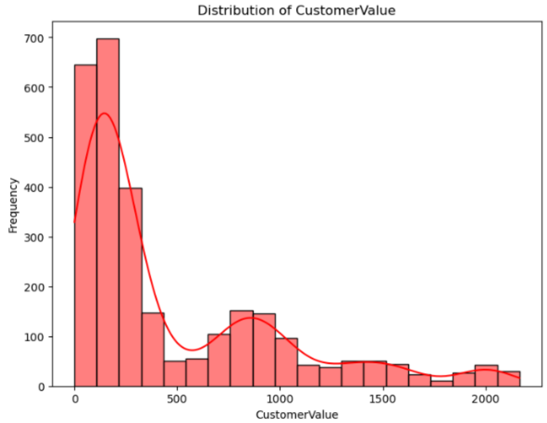
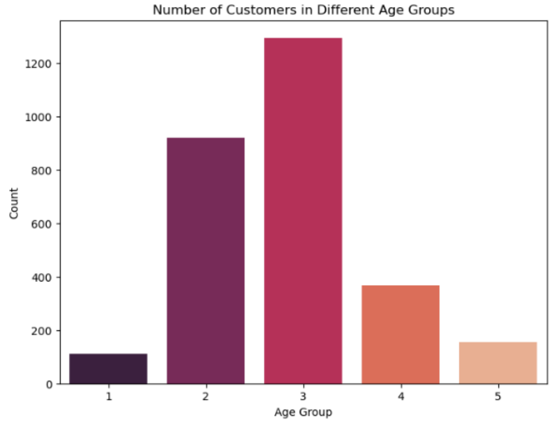
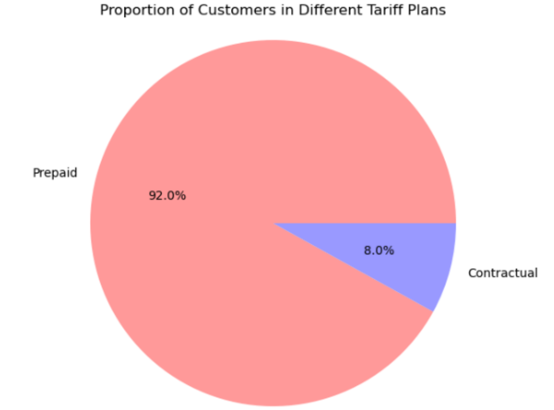
3.2 Task 2 - Data Exploration

3.2.1 Exploring each column

Chart	Observation	Key Takeaway

 *Figure 1. Distribution of Call Failures*	More than half of the customers reported experiencing fewer than 10 call features.	A low number of call failures suggested good network coverage, reducing the likelihood of technical issues for customers. However, the presence of extreme values (over 30) indicated customer dissatisfaction and a higher likelihood of churn.
 *Figure 2. Distribution of Subscription Length*	Most current customers remained loyal to the service for a duration of 30 to 40 months.	Most customers belonged to the loyal, long-term group, as opposed to the newly acquired customer segment.
 *Figure 3. Distribution of Frequency of Use*	The distribution was skewed to the left, with most customers having low levels of usage, as indicated by their total phone service usage falling under 100.	Despite customers having used the service for an extended period, the low range of total phone calls suggested infrequent use of the service.

 *Figure 4. Distribution of Seconds of Use*	The distribution showed a negative skew, with most values falling under 7500. The histogram's tail extended to the right, showing that some customers had higher usage, but these were fewer in number compared to most low-usage customers.	Customers demonstrated a pattern of short phone calls, although it remained more popular than SMS.
 *Figure 5. Distribution of Frequency of SMS*	The distribution exhibited a significant negative skew, with most values falling under 50. Extreme values were recorded up to 500.	Despite customers using the service for an extended period, the low range of total SMS indicated infrequent usage.
 *Figure 6. Distribution of Distinct Called Numbers*	The distribution exhibited a negative skew, with most values falling under 40.	Most values fell under 40, indicating that most customers have relatively low usage or engagement with the telecom services. Lower engagement could lead to higher churn rates, as these customers might find little value in continuing their subscriptions.

 *Figure 7. Distribution of Customer Value*	The distribution displayed a negative skew, with most values fall under 500.	Despite the findings from Subscription Length, most customers were estimated to fall within the lower range of value.
 *Figure 8. Number of Customers in Different Age Groups*	The most common age group fell between 25-30 (Age group 3), followed by 20-25 (Age group 2)	The current target customer group of this telecom company consisted mainly of young adults, who might tend to be less loyal to the service.
 *Figure 9. Proportion of Customers by Tariff Plans*	The dominant service segment was "pay-as-you-go", while contractual plans only accounted for 8%.	Customers on contractual plans may have a longer commitment to the service, making them less likely to churn compared to those on pay-as-you-go plans.

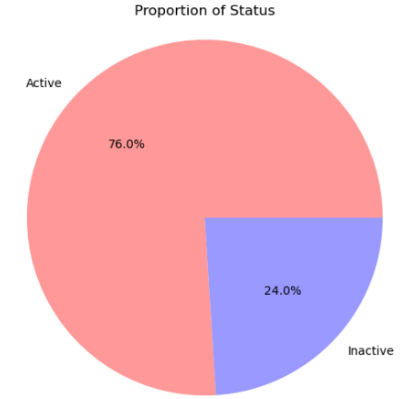
 *Figure 10. Proportion of Customers by Status*	Approximately less than a quarter of customers remained "inactive".	Although this is not a direct indicator of churn rate, active customers are less likely to churn compared to non-active customers.

Table 2 displays the data visualisation of individual column

3.2.2 Exploring relationships

Correlation matrix

Categorical attributes were converted into numerical to calculate correlation in this heatmap with the aim of selecting most relevant indicators of churn rate. Function ‘pd.factorize’ was applied to convert categorical columns and computes the correlation matrix of these numerical labels, which allowed the comparison among both numerical and categorical columns to ‘Churn’.

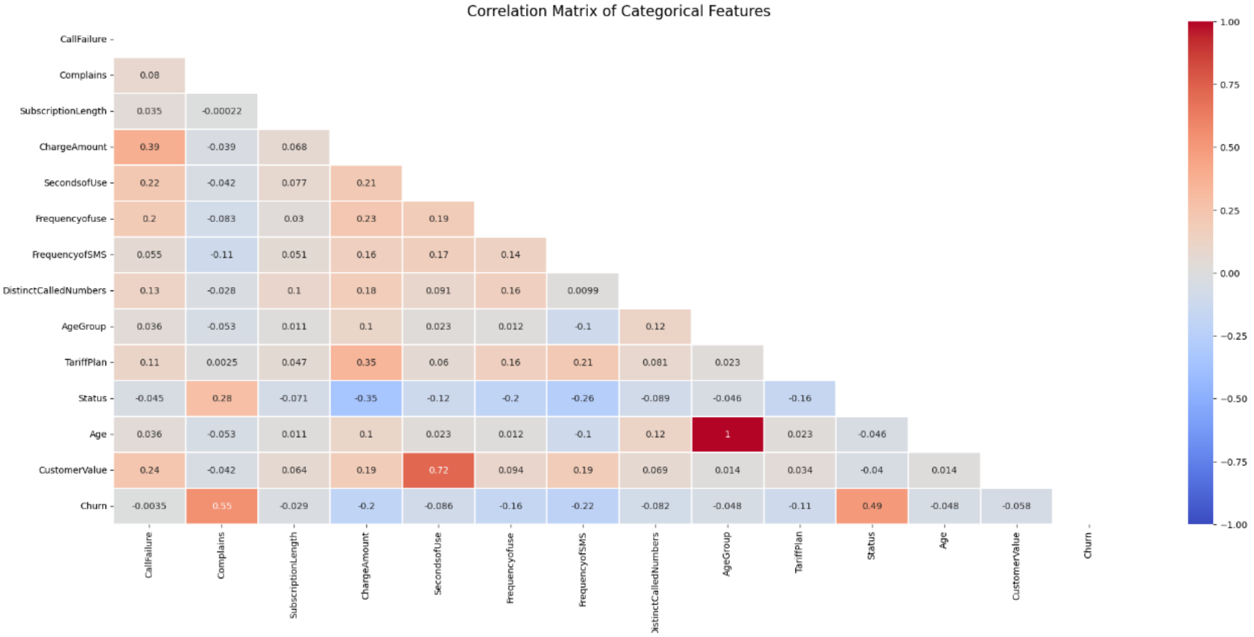


Figure 11. Correlation Heatmap of Numerical and Categorical Variables

- Positive correlation: Status and Complains were the two indicators with highest correlation with Churn.
- Negative correlation: Charge Amount and Status exhibited a slightly noticeable negative correlation, which indicated that as customers were charged more, they were more likely to abandon the service.

Key takeaway: Concerning positive correlations, customers who were active or inactive (Status) and customers who lodged complaints were more likely to churn. However, Charge Amount showed a negative correlation with Status, which implies that customers with higher charge amounts were less likely to be inactive.

Nested Pie Chart: Churn Distribution by Status

For plotting the relationship between Status and Churn, additional calculations were conducted to extract exact results for plotting. This involved calculating values for labels and sizes for churn using '.value_counts()' for two categories in the Churn column, and values for labels and sizes for Status by counting active and non-active statuses for two Churn categories. The 'explode' parameter is used to offset one or more slices of the pie chart.

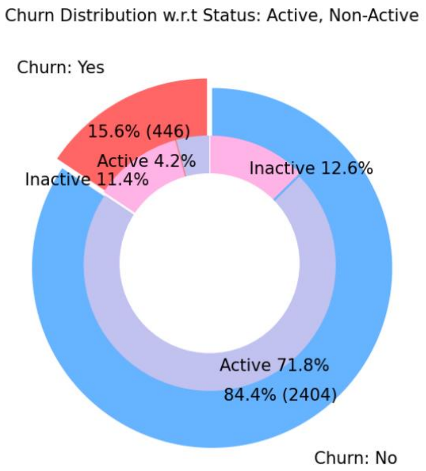


Figure 3: Observation
11.4% of inactive customers were observed to churn, while 71.8% of active customers did not churn.
However, 12.6% of active customers remained inactive, indicating a group with a higher risk of churn.

Figure 12. Distribution of Churn by Status

Key takeaway: Insights from churn analysis show that active customers are less likely to churn. However, the presence of inactive customers within the non-churn group suggests a risk of churn.

Stacked bar chart of Churn Distribution by Age Group

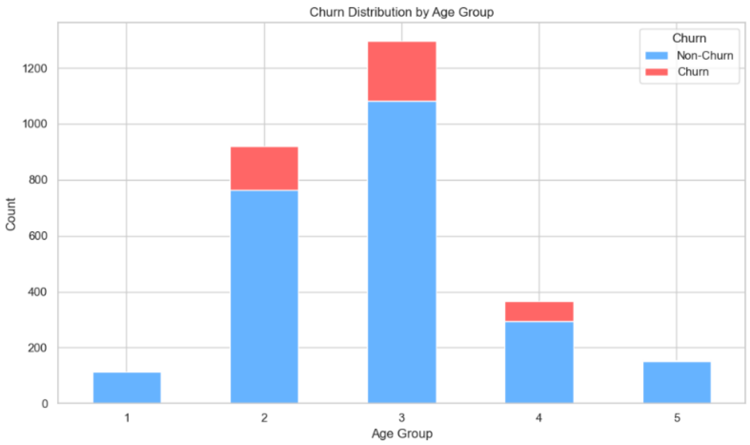


Figure 4: Observation
Churn customers are predominantly found in the 2nd and 3rd age groups. Interestingly, neither the youngest nor the oldest age group recorded any churn customers.

Figure 13. Distribution of Churn By Age Group

Key takeaway: The primary customer demographic for the telecom company consisted of young adults. This demographic tends to exhibit usage behaviours such as being less loyal, seeking new features and better deals, and being more likely to churn and switch providers. These findings align with previous observations, suggesting that younger age groups are more prone to churn.

Box plot of Churn Distribution by Frequency of Use

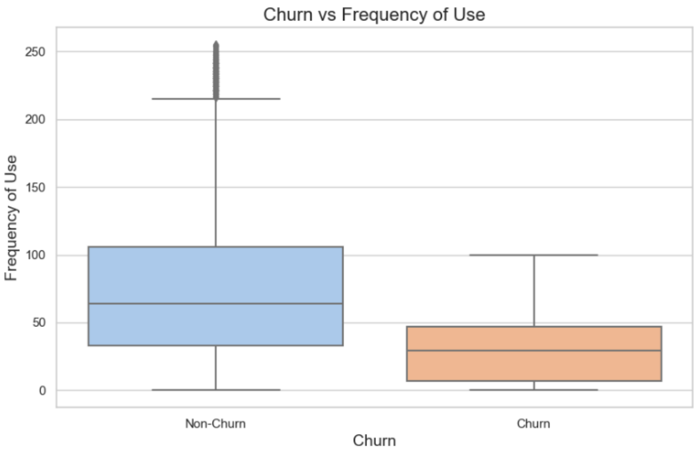


Figure 14. Distribution of Churn by Frequency of Use

Figure 4: Observation
The range of usage frequency for churn customers is notably lower compared to non-churn customers. Within the non-churn category, there are considerable outliers, suggesting challenges in predicting customers who are about to churn.

Key takeaway: Churn customers exhibited usage of under 100 calls. To reduce churn, the telecom company needed to encourage customers to use more than this threshold. Further research is necessary to investigate customers who were not churning but were approaching that point.

Bar chart of Churn Distribution by Tariff Plan

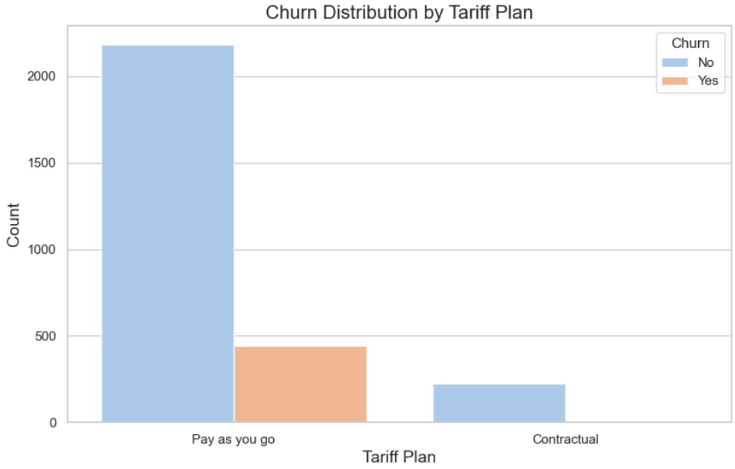


Figure 15. Distribution of Churn by Tariff Plan

Figure 4: Observation
The range of usage frequency for churn customers is notably lower than that for non-churn customers. There are significant outliers in the non-churn category, suggesting challenges in predicting customers who are about to churn.

Key takeaway: Prepaid customers showed a higher churn rate, whereas there was no churn in the contractual segment. The telecom company should focus on promoting customer loyalty, emphasizing the contractual segment as the more ideal service.

3.2.3. Posing one meaningful question and exploring the relationship

To develop a profile of customers at risk of churn, it is crucial to examine the financial incentives for retaining customers and determine whether it's worthwhile to retain them or acquire new ones. If retention is prioritized, a strategy should be devised.

Hypothesis Test

The hypothesis z-test was applied for proportions to see if there was a significant difference in churn rates between two groups of customers: those with higher customer value and those with low customer value.

The null hypothesis (H0) states that there is no difference in attrition rates between higher and lower value groups. Alternative hypothesis (H1): There is a difference in churn rates between high and low value groups. The null hypothesis was rejected because the p-value (2.30e-55) was substantially lower than the significance level (alpha = 0.05). Therefore, Churn rates differ significantly across high and low value groups.

Kernel Density Estimation (KDE) plot: Churn – Customer Value

KDE is a nonparametric approach for estimating the probability density function of a continuous random variable. Its flexibility is crucial in churn analysis (Mohammad et al., 2019) because of non-normal distributions in this dataset. The plot compares the distribution of Customer Value for churned customers (Churn == 1) and non-churned customers (Churn == 0).

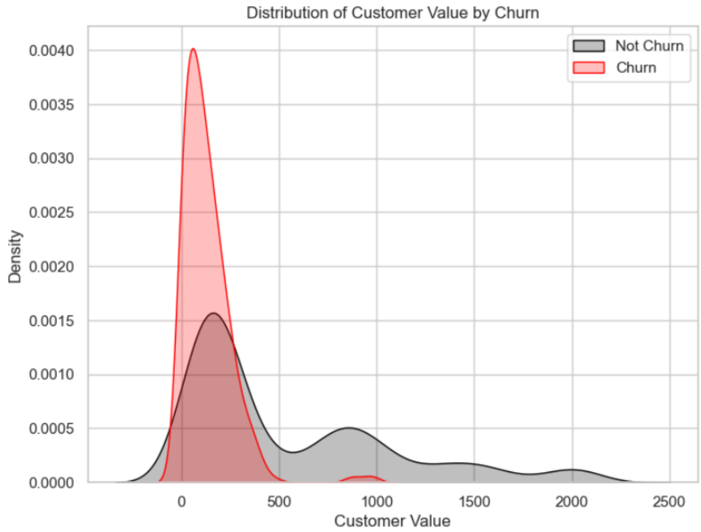


Figure 5: Observation
The red curve shifting to the left suggests a difference in the distribution of Customer Value between churned and non-churned customers.

Figure 16. Distribution of Customer Value by Churn

Key takeaway: Customers with higher customer value exhibit significantly different churn rates compared to those with lower customer value. This finding suggests that customer value is a crucial factor influencing churn. It indicates that customers with higher customer value are likely to be more loyal or satisfied, resulting in lower churn rates, whereas customers with lower customer value may churn at higher rates.

3.3. Task 3 - Data Modelling

3.3.1. Splitting the data into a training set and a test set

Data preprocessing

- Label Encoding: The function `object_to_int` is applied to convert categorical variables into integers using LabelEncoder. For this dataset, categorical columns are converted to binary and ordinal values already, however the data types were changed to 'object' for visualisation. There were no changes made to the values in dataset.
- Standardisation: Numerical columns are standardized using `StandardScaler` to ensure they have a mean of 0 and a standard deviation of 1, which can help algorithms converge faster and perform better.

Feature Selection

Geiler et al. (2022) implemented feature selections, such as SelectKBest, based on ANOVA F-value scores for churn analysis. Only the top five features are chosen: Complaints, Status, Frequency of Use, Seconds of Use, and Customer Value. These five columns correspond to the results of data investigation. Subsequently, the data with only the specified characteristics was divided into train and test sets.

Train-Test Data Splitting

For each suite, the dataset was divided into train and test sets using the `train_test_split` function from Scikit-Learn. The split was stratified based on the target variable 'Churn' to ensure a similar distribution of churn classes in both train and test sets. Only the selected top 5 features were used in the split to create X_train, X_test, y_train, and y_test for each suite.

Hyperparameter Tuning

Oversampling: Synthetic Minority Over-sampling Technique (SMOTE) was applied to address class imbalance, as the class distribution was heavily skewed, with only 495 instances of churn (class 1) compared to 2655 instances of non-churn (class 0).

SMOTE generates synthetic samples of the minority class (churned customers) by interpolating between existing instances in feature space. By creating new synthetic instances, SMOTE effectively augments the training data, allowing the model to learn more about the minority class and improve its ability to distinguish between churned and non-churned customers (Odusami et al., 2021).

Cross-Validation: GridSearchCV was used for hyperparameter tuning. In KNN's distance-based approach, hyperparameter tuning helped in mitigating the sensitivity to outliers, ensuring all features contribute equally to the distance computation.

3.3.2. Classification models

K-Nearest Neighbors (KNN)

From the observations in Figure 11, most columns show low correlation, and there's no distinct decision boundary. Utilizing KNN can be beneficial in such scenarios, as it's an instance-based learning algorithm. It classifies new instances by measuring their similarity to instances in the training set based on their features.

- Choose the number of neighbors k:

Parameter Definition: When n_neighbors is set to 1, the prediction is the class label of the nearest neighbor to the query instance. Increasing n_neighbors to higher values like 3, 5, 7, etc., allows the model to consider a larger number of neighbors, which helps in smoothing out decision boundaries and reducing overfitting.

Hyperparameter Tuning: Employing GridSearchCV involves exhaustively searching specified parameter values for an estimator. It requires the estimator (knn), parameter grid (param_grid), number of folds for cross-validation (cv=5), and the scoring metric ('accuracy' in this case).

Observations:

- The optimal k values for Suites 1 and 2 are both 3, indicating consistent model behavior even with different test sizes.

- The increase in the optimal k value for Suite 3 (k value = 5) suggests that slightly larger neighborhoods were preferred with smaller test sizes to improve generalization.

- Despite varying test set sizes, the model maintained relatively high accuracy across all suites, indicating robustness and generalization to unseen data.

- Train KNN model and make predictions:

KNeighborsClassifier was utilized to identify the k-nearest neighbors and assign labels based on majority vote among its k nearest neighbors. When the .fit() method was called on knn_model, the model was trained using the training data (X_train and y_train). The relationship between the features and the labels was learned by the model based on the provided training data.

Evaluation:

	Suite 1	Suite 2	Suite 3
Accuracy	0.900	0.908	0.916
Precision	0.647	0.661	0.688
Recall	0.798	0.843	0.843
F1 Score	0.715	0.741	0.758

Table 3 summarises the performance metrics of three suites by KNN

- Suite 3 generally outperformed the other suites in terms of accuracy, precision, recall, and F1 score. This suggests that the data in Suite 3 is more conducive to training a KNN model.

- Suite 1 and Suite 2 also perform reasonably well, but they have slightly lower accuracy and F1 score compared to Suite 3.

As recall and precision are important to churn analysis, Suite 2 was chosen for its balance between recall and precision.

Random Forest (RF)

A function called train_and_evaluate_rf was developed to accept training and testing data as input and train to evaluate the RF classifier. For the RF model, a grid of hyperparameters (grid_space) was constructed, which included maximum_depth, number of estimators, and criterion.

GridSearchCV was used to find the optimal hyperparameters (best_params_rf) by cross-validation. A new RF classifier (best_rf) was created using the best hyperparameters discovered.

Using the RF classifier on training data allows the model to discover patterns and correlations between features and the target variable. Following training, the model generates predictions on the test data, and performance parameters such as accuracy, precision, recall, and F1 score are measured. Additionally, a confusion matrix is generated to visualize the model's predictions compared to the actual labels.

Evaluation:

	Suite 1	Suite 2	Suite 3
Accuracy	0.890	0.914	0.907
Precision	0.620	0.715	0.691
Recall	0.776	0.747	0.730
F1 Score	0.689	0.731	0.710

Table 4 summarises the performance metrics of three suites by RF

Suite 2 generally outperforms the other suites in terms of accuracy, precision, recall, and F1-score. Suite 1 has the lowest performance in terms of precision and F1-score, while Suite 3 has the lowest recall.

Considering Suite 2's higher accuracy and balanced performance across precision, recall, and F1-score, its model might be the most robust and balanced among the three suites. Therefore, Suite 2 is chosen for its balanced performance in both recall and precision.

3.3.3. Comparing the two chosen models

Performance metrics

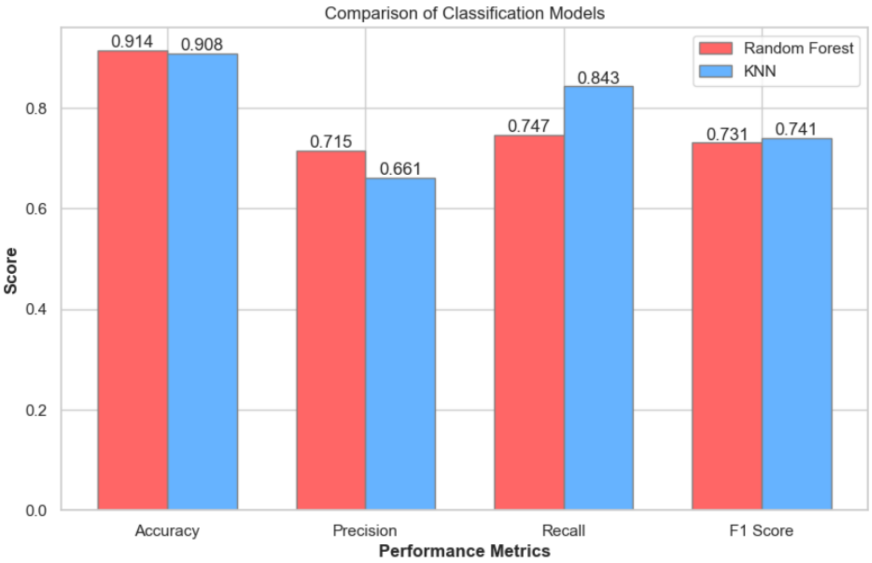


Figure 17 compares the performance metrics of KNN and RF models

- Accuracy: RF slightly outperformed KNN in terms of accuracy, indicating that it makes more correct predictions overall.
- Precision: RF had a higher precision (0.715), implying that RF was more efficient in churn customers identification without incorrectly designating too many non-churners as churners.
- Recall: KNN had a higher recall (0.843). In churn analysis, a higher recall is desirable as it means capturing as many actual churners as possible (Ullah et al., 2019).
- F1 Score: KNN had a slightly higher F1 score, suggesting that KNN achieved a better balance between precision and recall compared to RF.

KNN proved to be the better fit compared to RF due to its higher recall rate of 84.3% chances of correctly identifying actual churners among all the customers who were churning. While RF has somewhat higher accuracy and precision, these metrics reflect the overall soundness of predictions and the ability to prevent false positives. However, in the context of churn analysis, it is more important to collect as many genuine churn cases as possible, even if this results in some false positives. This is because failing to identify a churner may result in the loss of a customer, which is more serious than erroneously identifying a non-churner.

Confusion matrices

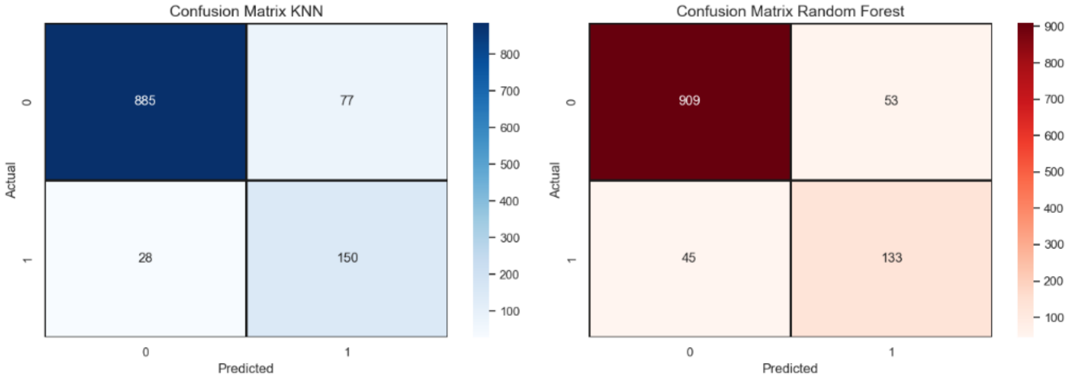


Figure 18 compares the confusion matrices of KNN and RF models

In the context of telecom churn analysis, missing a churn prediction (false negative) can result in revenue loss and customer unhappiness, whilst mistakenly identifying a non-churn customer as churned may result in wasteful retention efforts or even consumer displeasure. As a result, recall, which assesses the model's ability to recognize true churn cases, is critical in this situation.

Furthermore, KNN's instance-based learning technique classifies new instances based on their resemblance to examples in the training set (Wagh et al., 2024). KNN's ability to capture these relationships without making significant assumptions about the underlying data distribution was preferred for this customer churn dataset, as it contained unclear correlations between features.

4. Conclusions

In conclusion, the aim of this report is to compare classification models to predict and understand the characteristics of churn customers in the telecom industry. The analysis included standardization, feature selection using ANOVA, and hyperparameter tuning to enhance model performance and prevent overfitting. Among the train-test data pairs, Suite 2 exhibited the highest recall rate, which is crucial for accurate churn predictions. Considering the objective of capturing actual churn customers, KNN was chosen over the RF model due to its higher recall.

5. Reference

Geiler, L., Affeldt, S., & Nadif, M. (2022). A survey on machine learning methods for churn prediction. International Journal of Data Science and Analytics, 14(3), 217–242. https://doi.org/10.1007/s41060-022-00312-5

Mohammad, N. I., Ismail, S. A., Kama, N., Yusop, O. M., & Azmi, A. (2019). Customer churn prediction in telecommunication industry using machine learning classifiers. Proceedings of the 3rd International Conference on Vision, Image and Signal Processing. https://doi.org/10.1145/3387168.3387219

Odusami, M., Abayomi-Alli, O., Misra, S., Abayomi-Alli, A., & Sharma, M. M. (2021). A hybrid machine learning model for predicting customer churn in the telecommunication industry. In Advances in intelligent systems and computing (pp. 458–468). https://doi.org/10.1007/978-3-030-73603-3_43

Ullah, I., Raza, B., Malik, A. K., Imran, M., Islam, S. U., & Kim, S. W. (2019). A Churn Prediction Model using Random Forest: Analysis of machine learning techniques for churn prediction and factor identification in telecom sector. IEEE Access, 7, 60134–60149. https://doi.org/10.1109/access.2019.2914999

Wagh, S. K., Andhale, A. A., Wagh, K. S., Pansare, J. R., Ambadekar, S., & Gawande, S. H. (2024). Customer churn prediction in telecom sector using machine learning techniques. Results in Control and Optimization, 14, 100342. https://doi.org/10.1016/j.rico.2023.100342