Toronto Metropolitan University

The Chang School of Continuing Studies - Summer 2023

Course: CIND119 - Introduction to Big Data

Instructor: Davood Pirayesh Neghab, Ph.D.

Final Project:

Analysis of Decision Tree and Naïve Bayes Models for Customer Churn Prediction in Imbalanced Class Distribution

Group Members

Cheuk Chung Ho - cheuk.ho@torontomu.ca

Isolda Veruska de Almirante Silva - isolda.almirante@torontomu.ca

Athar Elahi - m1elahi@torontomu.ca

Table of Contents

1. Abstract.......... 2
2. Data Preparation.......... 2
2.1 Train-Validation-Test split.......... 3
2.2 One-Hot Encoding.......... 3
2.3 Performance of Decision Tree Before Major Data Processing.......... 4
2.4 Profile Report.......... 4
2.4.1 Imbalanced Class Distributions.......... 5
2.4.2 Feature with Only Distinct Values.......... 6
2.4.3 Correlation Between Class and Different Features.......... 7
2.4.4 Correlation Between “Number of Customs Service Calls” and “Churn”.......... 8
2.4.5 Correlation Between “Total Day Charge” and “Churn”.......... 9
2.4.6 Correlation Between “Int’l plan” “Mail Plan” and “Churn”.......... 10
2.5 Outliers from Different Features.......... 10
2.6 Elimination of the irrelevant and redundant features.......... 12
3. Predictive Modeling.......... 13
3.1 Baseline Modules.......... 13
3.2 Hyperparameters Tuning.......... 14
3.3 Threshold adjustment for the Naive Bayes Model.......... 16
3.4 Comparison Between Decision Tree and Naive Bayes After Adjustment.......... 17
3.5 Testing on The Test Set.......... 18
4. Conclusion and Recommendations.......... 19

1. Abstract

The primary objective of this project is to analyze and characterize customer churn and develop predictive models (Decision Tree and Naive Bayes) to determine whether existing customers will continue to use the telecommunications services of the company in the near future. Our team aims to gain valuable insights into customer behavior and patterns that contribute to churn from a dataset with 3,333 observations and 21 features, so as to help our company make better marketing strategies, optimize customer satisfaction, and enhance overall business performance.

To ensure the quality performance of the two classifiers that we developed, a variety of approaches were implemented throughout the project. In the exploratory data analysis stage, our team made some important findings, including the correlations between the labels and features, the correlation between different features, imbalanced class distributions, and the importance of the outliers. Baseline models were created after dropping the irrelevant features carefully. To further improve the models, our team fine-tuned the models with the best values of the hyperparameters and adjusted the threshold. With these approaches, our team concluded that the Decision Tree is a better model based on the metrics of recall for class true, precision for class true, and F1-score for the true class.

2. Data Preparation

The datasets have 3333 observations with below 21 features:

1. State: Customer's state.
2. Account Length: Integer number showing the duration of activity for customer account.
3. Area Code: Area code of customer.
4. Phone Number: Phone number of customer.
5. Inter Plan: Binary indicator showing whether the customer has international calling plan.
6. VoiceMail Plan: Indicator of voice mail plan
7. No of Vmail Mesgs: The number of voicemail messages
8. Total Day Min: The number of minutes the customer used the service during day time (continuous quantitative data type)
9. Total Day Calls: Discrete attribute indicating the total number of calls during day time.
10. Total Day Charge: Charges for using the service during day time (continuous data type).
11. Total Evening Min: The number of minutes the customer used the service during evening time.
12. Total Evening Calls: The number of calls during evening time.
13. Total Evening Charge: Charges for using the service during evening time.

14. Total Night Min: Number of minutes the customer used the service during night time.
15. Total Night Calls: The number of calls during night time.
16. Total Night Charge: Charges for using the service during night time.
17. Total Int Min: Number of minutes the customer used the service to make international calls.
18. Total Int Calls: The number of international calls.
19. Total Int Charge: Charges for international calls.
20. No of Calls Customer Service: The number of calls to customer support service.
21. Churn: Class attribute with binary values (True for churn and False for not churn)

2.1 Train-Validation-Test split

The dataset was split into three subsets: 60% for training, 20% for validation, and 20% for testing. This Train-Validation-Test split approach offers several advantages over a simple train-test split. Unlike a basic train-test split, this method provides a dedicated validation set that enables us to compare different model variations and fine-tune hyperparameters without touching the actual test set. This process significantly reduces the risk of overfitting and helps ensure that the final evaluation on the test set reflects the model's true generalization performance. As a result, it leads to a more reliable assessment of model effectiveness.

2.2 One-Hot Encoding

Among the 21 features in the dataset, 5 are categorical string variables: state, area code, phone number, International Plan, and VoiceMail Plan. Since both Decision Tree and Naive Bayes algorithms require numeric input, it was necessary to convert all string categories into numeric form. With all variables transformed into numeric types, the dataset can be efficiently processed by both algorithms. Additionally, having numeric data facilitates data visualization, as numeric categories can be easily represented using visual elements such as bars in histograms or points in scatter plots. There are two main approaches to converting string categories into numeric categories: label encoding and one-hot encoding. Since all of the categorical features in this dataset lack any inherent order or ranking, one-hot encoding was chosen for this project.

2.3 Performance of Decision Tree Before Major Data Processing

	precision	recall	f1-score	support
FALSE	0.95	0.98	0.96	577
TRUE	0.82	0.69	0.75	90
accuracy			0.94	667
macro avg	0.88	0.83	0.86	667
weighted avg	0.93	0.94	0.93	667

Figure 1. Output of Decision Tree Before data processing

In order to demonstrate that our data processing steps can improve the performance of the classifiers, our team trained a baseline Decision Tree model before applying any major data manipulation. As shown in the figure above, the Decision Tree's performance under limited preprocessing resulted in a precision of 0.82, a recall of 0.69, and an F1-score of 0.75 for the true class. This means that 82% of the instances predicted as true were actually correct, while 69% of the actual positive instances were correctly identified by the model. The F1-score of 0.75 indicates a reasonable trade-off between precision and recall.

2.4 Profile Report

From the Profile Report, it reveals several important findings:1) imbalanced class distributions, 2) Correlations between the target labels and different features, 3) correlations between different features, 4) the presence of outliers in some key features, 5) No missing values in the dataset

2.4.1 Imbalanced Class Distributions

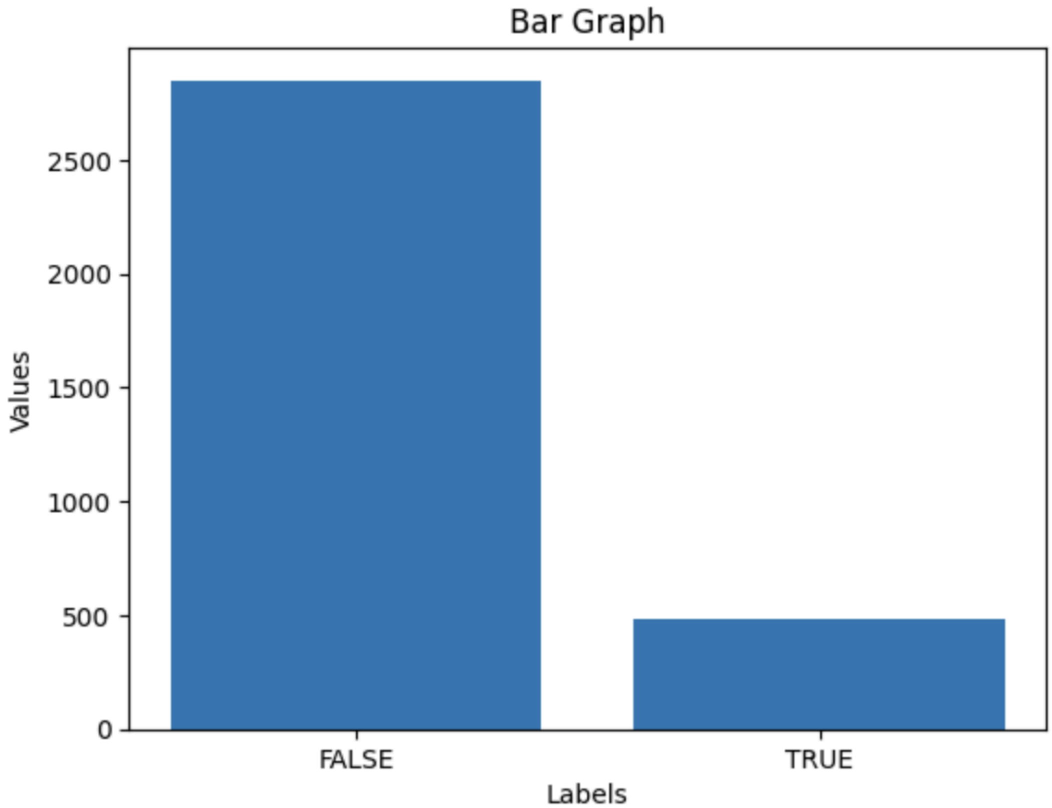


Figure 2. Bar Graph showing imbalanced class distribution

The findings from the report indicate a significant problem with imbalanced class distributions. 85.5% of the observations (2850) are classified as "False" (non-churn) meanwhile only 14.5% (483) as "True" (churn). with the "False" class being approximately 4.9 times more than the True" class. In this case, using accuracy as the key metric might be misleading as high accuracy can be achieved by a classifier that simply predicts the False (non-churn), but the classifier could be ineffective for predicting the minority class (churn).

Macro-averaged F1-score and weighted-averaged F1-score are often considered good metrics for evaluating performance on imbalanced class distributions. However, in our case, the primary objective is to identify customer behaviors that lead to churn, which corresponds to the minority class ("True") in the imbalanced distribution. As a result, assigning weights based on sample size or averaging across classes could potentially misrepresent the model's effectiveness in identifying churn instances, thereby weakening the insights we aim to extract from the dataset.

Given the imbalanced distribution and our focus on the minority class, the F1-score for the "True" class will be the key performance metric used to evaluate classifiers in this project. This metric is particularly valuable when working with imbalanced datasets where the minority class is of greater interest, as it incorporates both precision and recall. It penalizes classifiers that have a high number of false negatives (low recall) or false positives (low precision), balancing the trade-off between false positives and false negatives.

2.4.2 Feature with Only Distinct Values

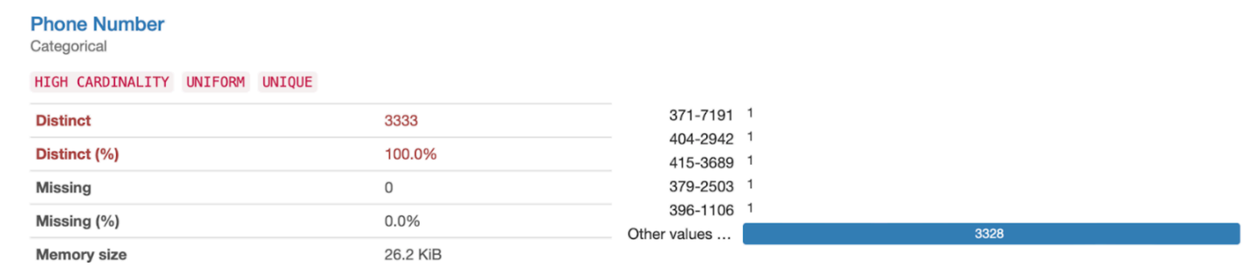


Figure 3. Bar Chart for phone number and its description

Of all the features, there is one feature that is only with distinct values, which is the “Phone Number.” As expected, all 3,333 observations have different phone numbers. This means the “Phone Number” feature cannot influence customers’ behavior and the performance of the model. Therefore, this feature can be dropped and not be considered for further analysis.

2.4.3 Correlations Between Class and Different Features

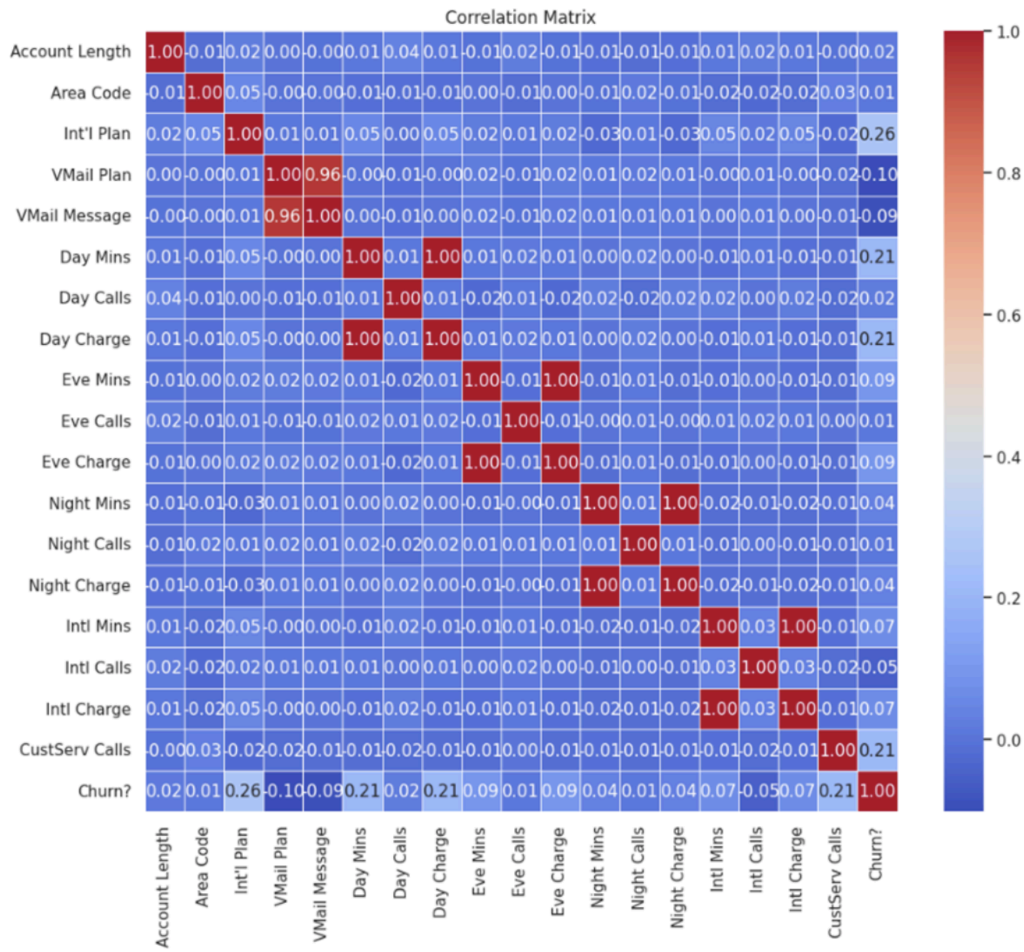


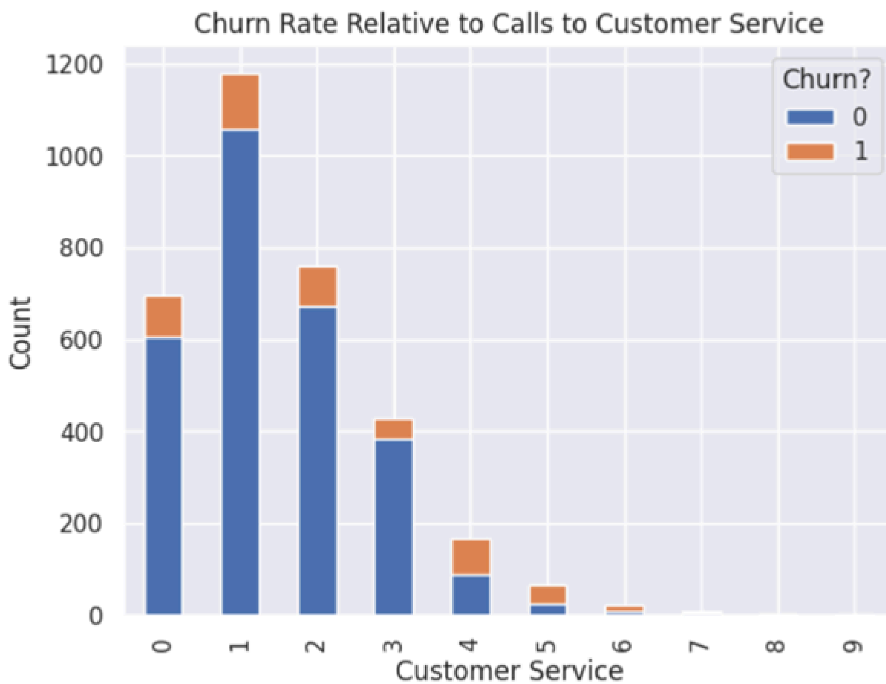
Figure 4. Correlation Matrix between the class (Churn) and different features

From the correlation matrix graph above, it can be found that the class has a significant correlation with 4 features, including "Total Day Min", "Total Day Charge", "No of Calls Customers Service", and "Inter Plan". As for “Account Length”, “Phone Number (unique)”, “Total Evening Calls”, “Area Code”, “Total Night Calls”, “Total Night Minutes”, “Total Night Charge”, and “Total Day Calls”, these features have low or no correlation with the class "churn.

Another observation from the matrix is the correlation between the “charges” and “minutes”, including “Total Day Charge” and “Total Day Min”, “Total Evening Charge” and “Total Evening Min”, “Total Night Charge” and “Total Night Minutes”, and “ Total Int

Charge” and “Total Int Min”. This is expected as they are highly correlated or even 100% correlated since the charges are directly proportional to the minutes of call usage.

2.4.4 Correlation Between “Number of Customs Service Calls” and “Churn”



Churn? Int'l Plan	0	1
0	2664	346
1	186	137

Figure 5. Stacked Bar of Customer Service and Churn and proportions

The graph below compares the customer churn rate with the number of calls to customer service. It shows that customers who contact customer service more frequently are more likely to cancel or abandon the company's services compared to other customers.

It can be observed that there is a significant increase in the number of churn cases among customers who make four or more calls to customer service. This suggests that the problems customers experience with the service, and whether their issues or concerns are adequately resolved, are important factors influencing their decision to churn. If customers encounter too many issues and are not satisfied with the support they receive, it often leads to their decision to leave and seek another company with better service.

2.4.5 Correlation Between “Total Day Charge” and “Churn”

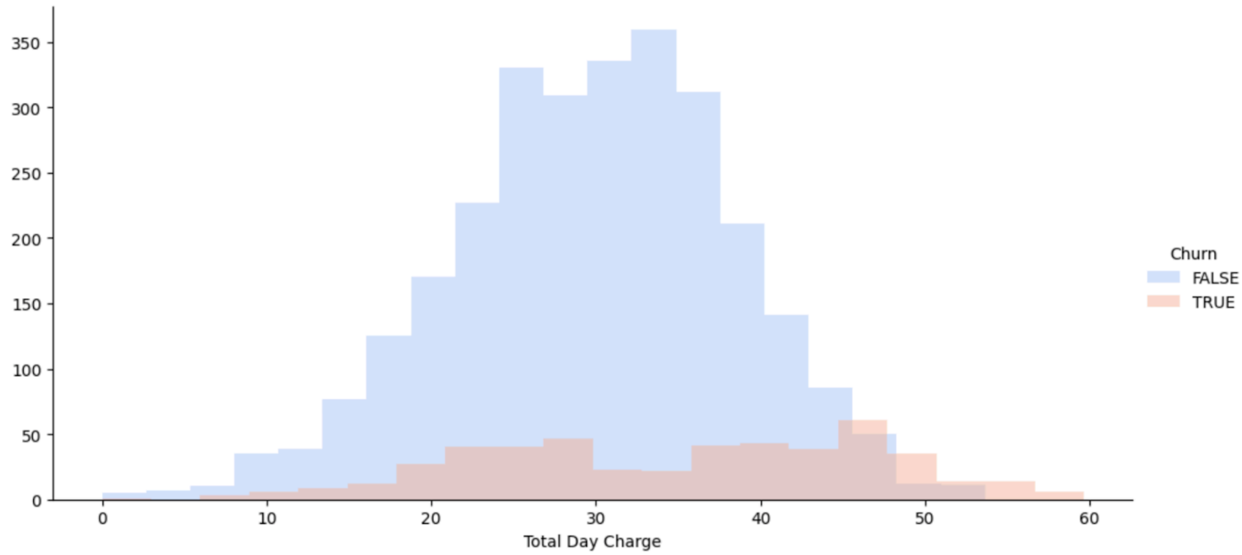


Figure 6. Histogram showing correlation between Total Day Charge and Churn

From the chart above, it can be observed that while Total Day Charge is normally distributed, churn as a function of Total Day Charge is left-skewed, suggesting a possible association between higher day charges and increased likelihood of churn. When analyzing the data at the customer level, it was found that the majority of customers with more than 45 in Total Day Charges had churned, and only customers who churned paid more than 55 in charges. In contrast, none of the customers who paid the lowest charges had churned. This suggests that price levels are a factor when customers are deciding between staying or switching phone companies.

2.4.6 Correlation Between “Int’l plan” “Mail Plan” and “Churn”

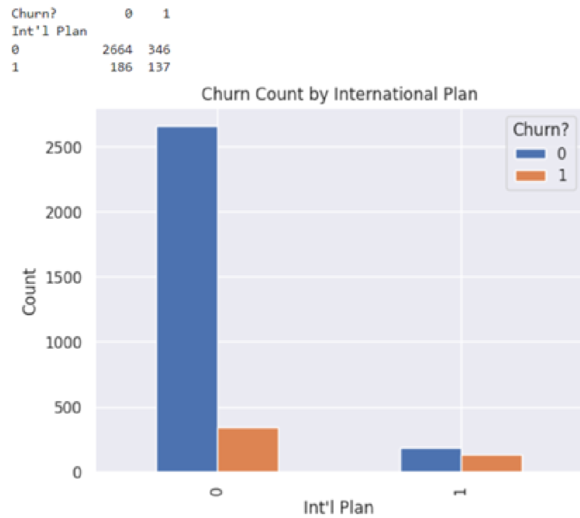


Figure7. Bar graphs showing correlation of Churn with International Plan

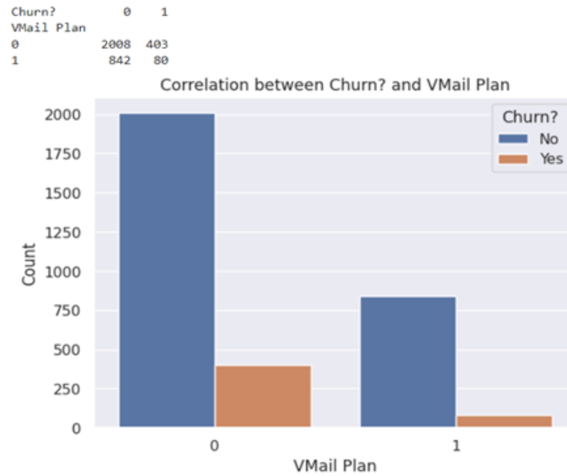


Figure 8. Bar graphs showing correlation of Churn with VoiceMail.

The cross-tabulation reveals that customers with the International Plan are less likely to churn (137 churned out of 323) compared to those without the plan (346 churned out of 3,003). The presence of the International Plan appears to have a positive effect on customer retention, indicating that offering additional features like the international plan may reduce churn rates and enhance customer loyalty. The relationship between customers having a VMail Plan and their churn status shows that customers without a VMail Plan are more likely to churn (403 out of 2,411) compared to those with a VMail Plan (80 out of 922). This suggests that having a VMail Plan may play a role in reducing churn rates and retaining customers.

2.5 Outliers from Different Features

	Account Length	No of Vmail Mesgs	Total Day Min	Total Day calls	Total Day Charge	Total Evening Min	Total Evening Calls	Total Evening Charge	Total Night Minutes	Total Night Calls	Total Night Charge	Total Int Min	Total Int Calls	Total Int Charge	No of Calls Customer Service
count	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000	3333.000000
mean	101.064806	8.099010	179.775098	100.435644	30.562307	200.980348	100.114311	17.083540	200.872037	100.107711	9.039325	10.237294	4.479448	2.764581	1.562856
std	39.822106	13.688365	54.467389	20.069084	9.259435	50.713844	19.922625	4.310668	50.573847	19.568609	2.275873	2.791840	2.461214	0.753773	1.315491
min	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	23.200000	33.000000	1.040000	0.000000	0.000000	0.000000	0.000000
25%	74.000000	0.000000	143.700000	87.000000	24.430000	166.600000	87.000000	14.160000	167.000000	87.000000	7.520000	8.500000	3.000000	2.300000	1.000000
50%	101.000000	0.000000	179.400000	101.000000	30.500000	201.400000	100.000000	17.120000	201.200000	100.000000	9.050000	10.300000	4.000000	2.780000	1.000000
75%	127.000000	20.000000	216.400000	114.000000	36.790000	235.300000	114.000000	20.000000	235.300000	113.000000	10.590000	12.100000	6.000000	3.270000	2.000000
max	243.000000	51.000000	350.800000	165.000000	59.640000	363.700000	170.000000	30.910000	395.000000	175.000000	17.770000	20.000000	20.000000	5.400000	9.000000

Figure 9. Table with descriptive statistics of the variables

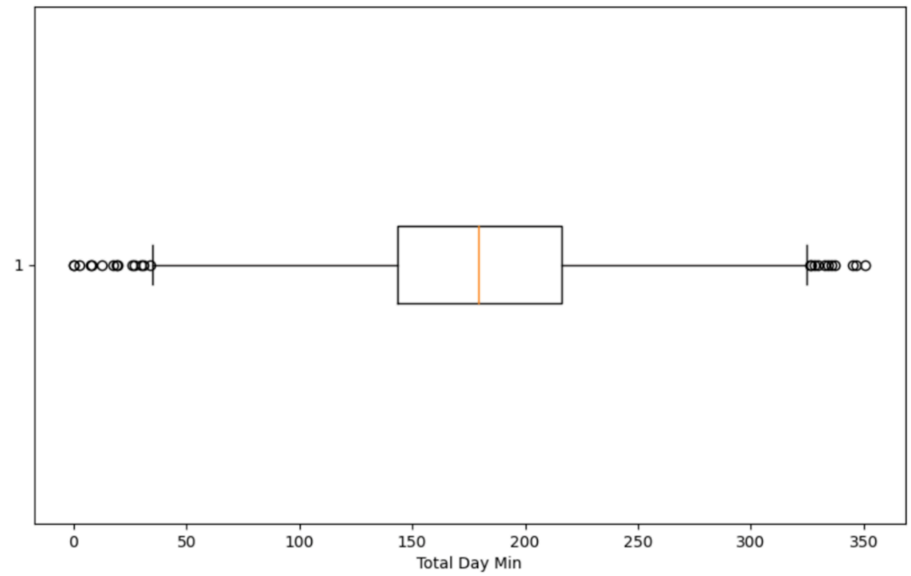


Figure 10. Box Plot showing outliers of the Total Day Minutes variable

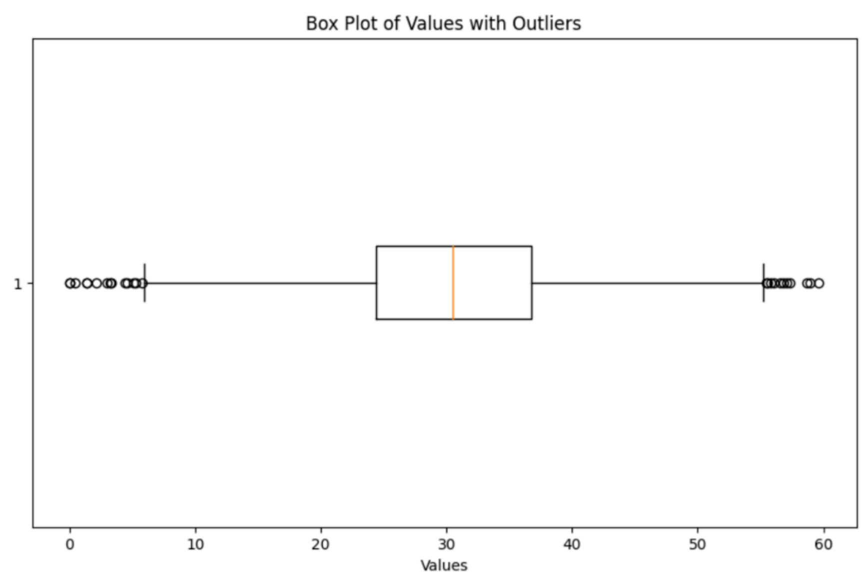


Figure 11. Box Plot showing Values with outliers

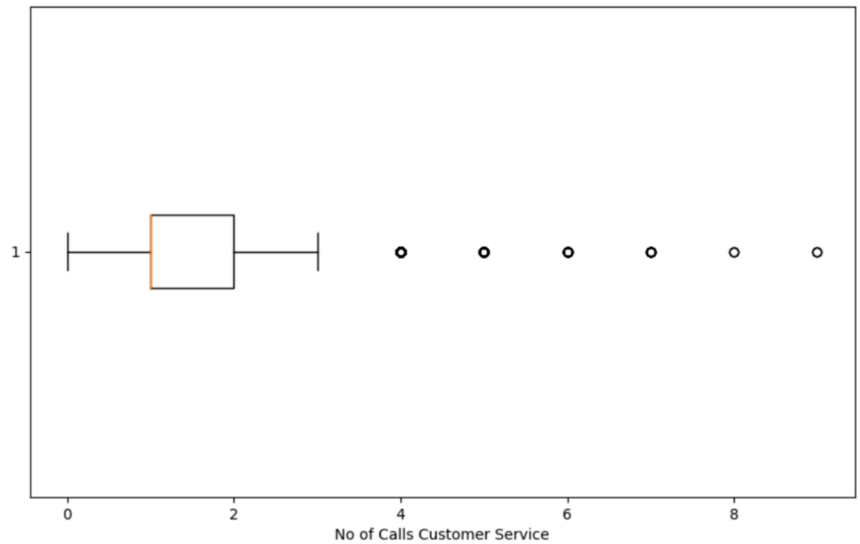


Figure 12. Box Plot showing outliers of the Customer Services Calls

From the table and boxplots above, it clearly shows that there are some outliers in the dataset, but some valuable insights can be found because of the outliers. Taking the correlation between “churn” and “total day charge” and “churn” and “number of calls to customer service” as examples, customers who made an excessive number of calls to customer service have churned and those who paid an excessive amount of charges have also churned. These two examples demonstrate the importance of the outliers in this dataset.

Considering the importance of the outliers and the fact that there is no missing value in the dataset, no major adjustment is needed for our dataset, except for removing some irrelevant attributes.

2.6 Elimination of the Irrelevant and Redundant Features

Based on the exploratory data analysis, our team decided to drop some of the attributes, including “Total Day Min”, “Total Evening Min”, “Total Int Min”, “Account Length”, “Total Night Calls”, “Total Night Minutes”, “Total Night Charge”, “Total Evening Calls”, “Area Code”, “Phone Number”, “Total Day calls”

“Account Length”, “Phone Number”, “Total Evening Calls”, “ Area Code”, “Total Night Calls”, and “Total Night Minutes”, and “Total Night charge” were dropped because their correlation with the class is significantly low. These features do not influence customers’ decisions and thus are not suitable for training classifiers that predict whether customers will churn.

“Total Day Min”, “Total Evening Min”, “Total Int Min”, and “Total Night Minutes” are considered unnecessary to keep. As mentioned above, all the “Minutes” features are significantly correlated, or even 100% correlated with the corresponding “Charges”, making them redundant for the models.

3. Predictive Modeling

3.1 Baseline Modules

After eliminating irrelevant and redundant attributes, our team developed two baseline models for the churn prediction task: a Decision Tree classifier and a Naive Bayes classifier. These models serve as the initial starting point for further refinement and evaluation. To ensure that the attributes we dropped were truly irrelevant and redundant, we compared the performance of the Decision Tree with its initial results (before major data processing) and confirmed that there was no substantial change. The figures below present the performance of the baseline models across various evaluation metrics. The first figure (left) shows the results for the Decision Tree classifier, while the second figure (right) shows the results for the Naive Bayes classifier.

	precision	recall	f1-score	support
FALSE	0.95	0.99	0.97	564
TRUE	0.93	0.72	0.81	103
accuracy			0.95	667
macro avg	0.94	0.85	0.89	667
weighted avg	0.95	0.95	0.94	667

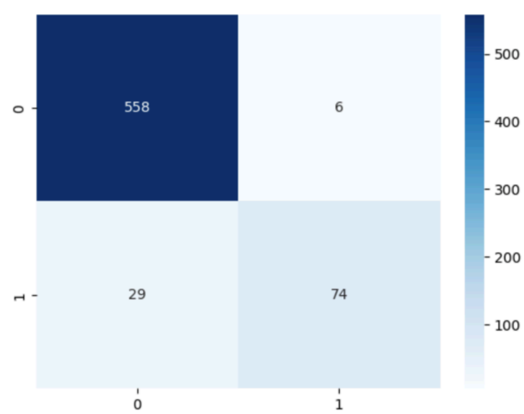


Figure 13. Baseline for Decision Tree on different Metrics

	precision	recall	f1-score	support
FALSE	0.91	0.36	0.52	564
TRUE	0.19	0.80	0.30	103
accuracy			0.43	667
macro avg	0.55	0.58	0.41	667
weighted avg	0.80	0.43	0.48	667

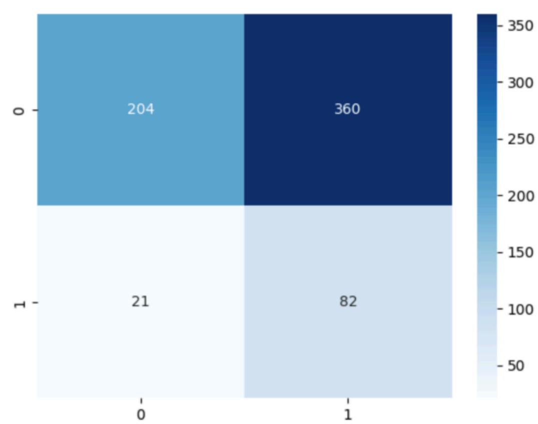


Figure14. Baseline for Naive Bayes on different Metrics

The results clearly show that the Decision Tree model generally outperforms Naive Bayes in terms of overall predictive performance. The Decision Tree achieves a precision of 0.93, a recall of 0.72, and an F1-score of 0.81 for the true class (churn), while Naive Bayes records a precision of only 0.19, a recall of 0.80, and an F1-score of 0.30 for the same class.

Although the recall of Naive Bayes is slightly higher than that of the Decision Tree by 0.08, its much lower precision results in a significant number of false positives. Specifically, a large number of non-churned customers (360 observations) are incorrectly classified as churned. This leads to a low F1-score due to the imbalance between precision and recall.

These results demonstrate that the Decision Tree is better at correctly identifying true churned customers while minimizing false alarms. Therefore, the Decision Tree is considered to be a more effective and reliable model at this stage.

3.2 Hyperparameters Tuning

Given that each model has many hyperparameters, it is difficult to manually determine the optimal values for each. To address this, our team performed a grid search to identify the best combination of hyperparameters for the models. The best values of hyperparameters for the decision Tree classifier are as follows: “criteria” = entropy, “max_dept” = 10, “min_samples_leaf” = 2, “min_samples_spli” = 4

	precision	recall	f1-score	support
FALSE	0.96	0.98	0.97	564
TRUE	0.89	0.80	0.84	103
accuracy			0.95	667
macro avg	0.93	0.89	0.91	667
weighted avg	0.95	0.95	0.95	667

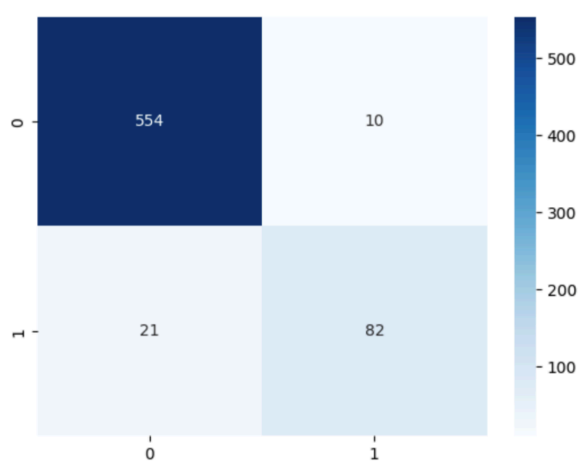


Figure 15. Baseline for Decision Tree on different metrics after hyperparameter tuning

After tuning the hyperparameters for the Decision Tree model, there was a noticeable improvement in performance. The recall for the true class (churn) increased by 0.08, rising from 0.72 to 0.80. Out of 103 actual churn cases, the model is now able to correctly identify 82 of them. This means the fine-tuned Decision Tree captures 8% more of the actual churning customers compared to the baseline model.

Although precision slightly decreased by 0.04, from 0.93 to 0.89, this trade-off is considered acceptable. In this context, correctly identifying customers who are likely to churn is more critical than avoiding false positives, as the primary goal is to understand why customers are churning in order to address the underlying issues. The increase in the F1-score, from 0.81 to 0.84, demonstrates that the model has achieved a better balance between precision and recall, further justifying the trade-off.

	precision	recall	f1-score	support
FALSE	0.91	0.37	0.53	564
TRUE	0.19	0.80	0.30	103
accuracy			0.44	667
macro avg	0.55	0.59	0.42	667
weighted avg	0.80	0.44	0.50	667

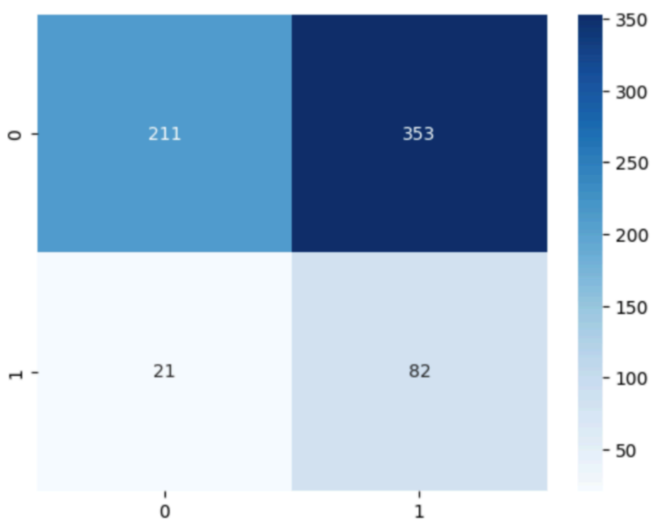


Figure 16: Naive Bayes on different metrics after hyperparameters tuning

As for the Naive Bayes model, fine-tuning the hyperparameters did not lead to any significant improvement. All key metrics—recall, precision, and F1-score—remained unchanged. Therefore, we decided to explore additional ways to enhance the model's performance. After identifying the optimal classification threshold, our team adjusted the model accordingly for further improvement.

3.3 Threshold adjustment for the Naive Bayes Model

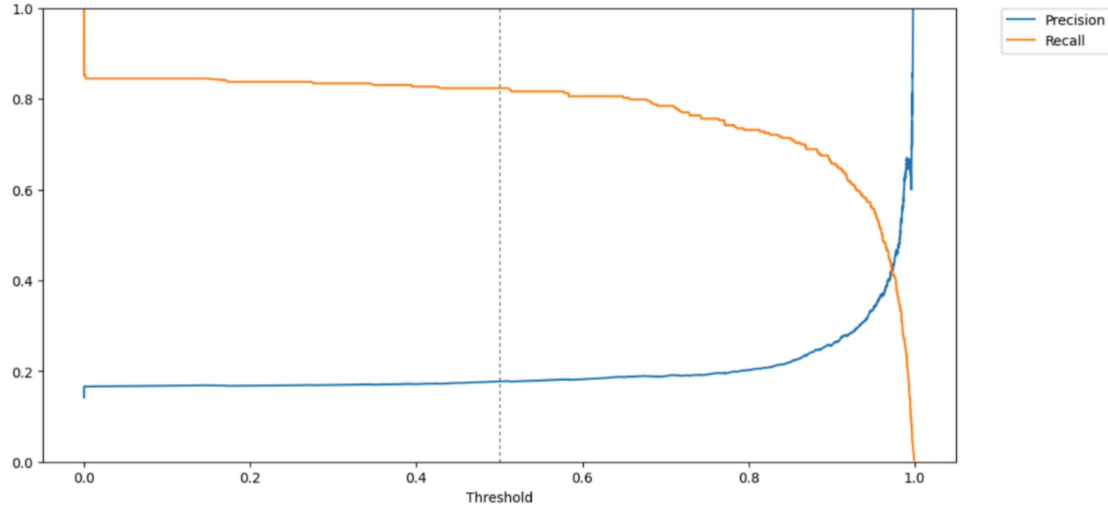


Figure 17: Precision and Recall after Threshold

	precision	recall	f1-score	support
FALSE	0.92	0.83	0.87	564
TRUE	0.38	0.58	0.46	103
accuracy			0.79	667
macro avg	0.65	0.70	0.66	667
weighted avg	0.83	0.79	0.81	667

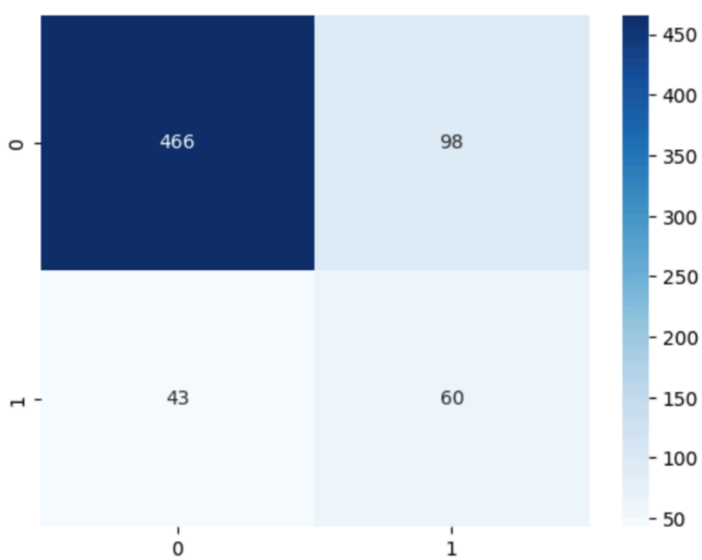


Figure 18: Naive Bayes on different metrics after threshold adjustment

Based on the plot above, we observed that precision can increase significantly at the expense of recall. After conducting multiple trials with different thresholds, we found that a threshold of 0.94 produced the most favorable results.

At this threshold, there is a noticeable drop in recall for the true class (churn), decreasing from 0.80 to 0.58. The model now correctly predicts 60 out of 103 actual churn cases (58%). However, this adjustment leads to a substantial improvement in both precision and F1-score for the churn class.

Previously, the precision for the churn class was only 0.19—meaning that out of 434 predictions labeled as churn, only 82 were correct. After adjusting the threshold to 0.94, precision increases by 200% to 0.38, with 62 correct predictions out of 158 predicted churn cases.

While correctly identifying churned customers is the primary focus of this task, the model must also avoid completely neglecting the majority class (non-churn). As discussed earlier, the F1-score is the key evaluation metric in this project because it balances both precision and recall. It penalizes classifiers for both false positives (low precision) and false negatives (low recall), providing a more balanced view of model performance.

The increase in F1-score from 0.30 to 0.46 demonstrates that our threshold adjustment effectively improved the Naive Bayes model's performance.

3.4 Comparison Between Decision Tree and Naive Bayes After Adjustment

After the adjustments, both the Decision Tree and Naive Bayes models show improved overall performance. The Decision Tree now achieves a precision of 0.89, a recall of 0.80, and an F1-score of 0.84. In comparison, the Naive Bayes model reaches a precision of 0.38, a recall of 0.58, and an F1-score of 0.46.

Although fine-tuning and threshold adjustment enhanced Naive Bayes' performance, it still exhibits significantly lower precision, recall, and F1-score for the true class (churn). This suggests that Naive Bayes may struggle to accurately identify churned customers, potentially leading to missed retention opportunities and reduced overall effectiveness.

Therefore, the Decision Tree remains the superior model at this stage. It demonstrates a stronger ability to identify churning customers by capturing a larger proportion of actual churn cases while maintaining a better balance between recall and precision.

3.5 Testing on The Test Set

	classifier	threshold	precision	recall	f1-score
0	Decision Tree	0.50	0.83	0.74	0.78
0	Naive Bayes	0.94	0.32	0.59	0.41

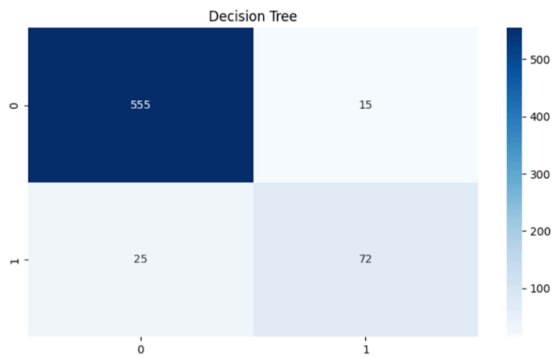


Figure 19. Baseline for Decision Tree

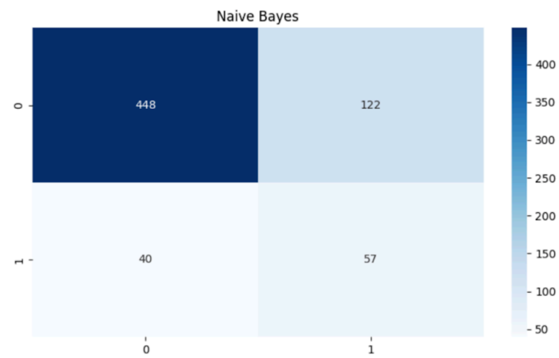


Figure20. Baseline for Naive Bayes

On the test set, both classifiers perform worse compared to their performance on the validation set, which is expected since the models were fine-tuned on the validation data.

For the Decision Tree, the precision, recall, and F1-score for the true class (churn) decrease to 0.83, 0.74, and 0.78, respectively. In the case of Naive Bayes, the precision for the true class drops from 0.38 to 0.32, while recall slightly increases from 0.58 to 0.59. However, since the F1-score balances precision and recall and is the key metric for this project, its decrease from 0.46 to 0.41 indicates an overall weaker performance of Naive Bayes on the test set.

Despite the decline in the Decision Tree's performance on the test set compared to the validation set, it still noticeably outperforms Naive Bayes across all metrics—precision, recall, and F1-score for the true class. This means the Decision Tree remains better at predicting customers who churn, capturing a larger proportion of actual churn cases, and maintaining a better balance between precision and recall.

4. Conclusion and Recommendations

Our group finalized the Decision Tree and Naive Bayes models after data processing, hyperparameter tuning, and threshold adjustment. Given that the Decision Tree consistently achieves significantly higher recall, precision, and F1-score for the true class (churn) than Naive Bayes on both the validation and test sets, it is clear that the Decision Tree is the superior classifier for this task.

Naive Bayes performs poorly on this dataset largely due to its assumption of feature independence and sensitivity to rare events, which negatively impacts its accuracy in estimating probabilities for the minority class. In contrast, the Decision Tree is a non-parametric model that makes no strong assumptions about the underlying data distribution. This flexibility allows it to better handle imbalanced data by creating splits that isolate minority class instances, resulting in improved classification performance. Therefore, we recommend using the Decision Tree model to identify potential churners with the goal of implementing targeted marketing strategies to enhance customer satisfaction and overall business performance.

Through predictive modeling and data analysis, we uncovered valuable insights to reduce churn rates and boost customer retention. Notably, customers with International Plans and Voicemail Plans exhibit lower churn rates, indicating that offering such supplementary services can significantly increase customer loyalty. Promoting these plans and other attractive features may help reduce customers' likelihood of switching to competitors.

We also observed that customers who churn tend to make numerous calls to customer service centers, with more than four calls emerging as a critical threshold linked to frustration and eventual churn. Providing excellent customer service is therefore essential to prevent churn. Enhanced training for customer service agents can improve their ability to effectively resolve customer issues, contributing to higher satisfaction and retention.

Additionally, offering discounts to customers facing high charges could serve as an effective retention strategy. Identifying these customers and providing targeted discounts demonstrates the company's commitment to meeting their needs and concerns, fostering loyalty.

In conclusion, focusing on additional service offerings, investing in customer service training, and revising service charges with targeted discounts are key strategies to improve customer retention and reduce churn. Furthermore, gathering direct customer feedback is vital for understanding the root causes of churn and making informed decisions to enhance overall customer satisfaction and loyalty.

Workload Distribution

Member Name	List of Tasks Performed
Cheuk Chung Ho	Abstract, Data preparation, Predictive Modeling, and Conclusions and Recommendation
Isolda Veruska de Almirante Silva	Data preparation, and Conclusions and Recommendation
Athar Elahi	