Predicting Customer Churn

Project Report

Raphael Rivers

August 25, 2024

Abstract

The primary objective of this project is to predict customer churn and identify the key features influencing this behavior using the customer service call dataset provided by PPG. The dataset does not contain personally identifiable information and consists of data points highlighting customer characteristics and company interactions. This study has developed robust predictive models through a systematic approach involving data preprocessing, exploratory data analysis (EDA), feature engineering, model building, model performance evaluation, and cross-validation. The analysis identified the 2 main variables, namely, 'total_day_minutes,' 'total_eve_minutes,' and a principal component feature, 'pca_cluster,' as significant predictors of churn. This project yielded two models that can be used for feature prediction on a new dataset. The primary model, Model 8, included interaction terms and polynomial features and demonstrated the highest performance with a mean ROC_AUC score of 0.6798 and a 95% confidence interval of 0.0117. These findings underscore the importance of advanced feature engineering and model selection techniques in accurately predicting customer churn. More importantly, they provide actionable insights for enhancing customer retention strategies, thereby contributing to the stability and growth of businesses.

Introduction

Customer churn is a phenomenon where customers discontinue their subscription to a service, which is a critical challenge faced by companies, particularly in the product manufacturing and services industries (Luck, 2023). Therefore, predicting the likelihood of one or more customers discontinuing their purchases enables a company to proactively implement strategies aimed at satisfying and retaining those customers. Understanding and predicting churn behavior is essential for developing effective retention strategies and maintaining a stable customer base (Smolic, 2023). Most importantly, the ability to identify customers who are likely to churn allows businesses to take proactive measures to enhance customer satisfaction and loyalty. This project aimed to predict customer churn using logistic regression models, including both linear and interaction terms, and to validate these models using cross-validation. This churn prediction project involves data preprocessing, exploratory data analysis, feature engineering cluster analysis, principal component analysis, statistical model building, interpreting, prediction, model performance validation, and evaluation.

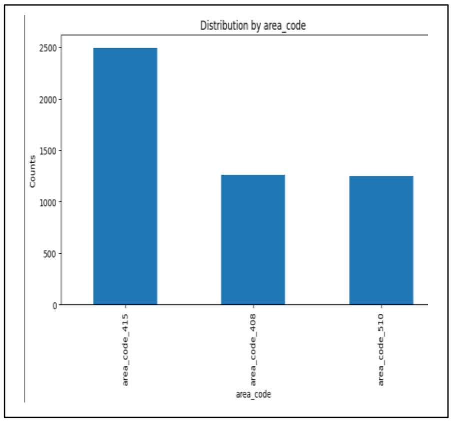
Data Preprocessing

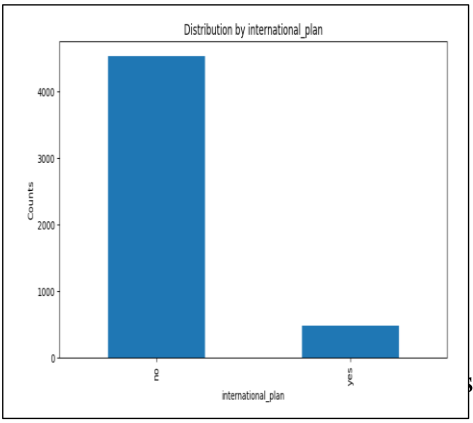
Cleaning and transforming the data to prepare it for analysis involved initially loading the dataset and thoroughly inspecting it to identify any missing values, variable types, and distributions of the variables. Key preprocessing steps included identifying skewed variables for log-transforming to reduce skewness and standardizing numerical features to ensure they have a mean of zero and a standard deviation of one (Akhtar, 2023). This comprehensive data preprocessing and transformation process is essential to prepare the dataset for subsequent analysis and modeling.

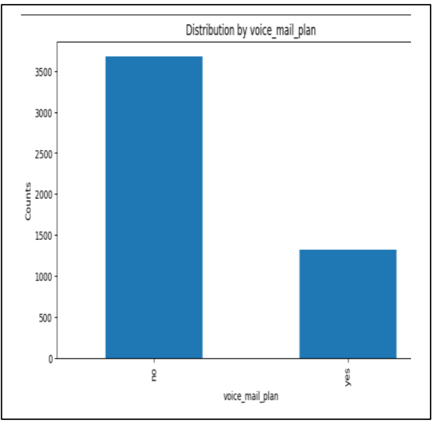
Exploratory Data Analysis (EDA)

The initial exploration of the dataset reveals important insights into the data, customer behavior, and variable relationships. The dataset comprises 5000 rows and 20 columns *(Appendix A)*. The input variable(s), or features, consists of 19 columns that contain data about customer characteristics, usage, plans, and interactions, likely influencing whether a customer will churn. The total number of unique values across all variables comes to 10239. The categorical variables include 'area_code' with 3 unique values. Comprising of Area codes 415 with 1655, 408 with 1655, and 510 with 1690. Other categorical variables such as 'international_plan' has 2 unique values of events (no: 4527, yes: 473), and 'voice_mail_plan' shows 2 unique event values (no: 3678, yes: 1322), and churn the output variable contains 2 unique event values (false: 4293, true: 707) *(Appendix A)*. Data visualization revealed a strong correlation between usage and charges and the impact of different plans on the churn variable *(Appendix B)*. This variable is crucial in building predictive models for customer churn. The dataset is clean with no missing values, and the categorical variables have sufficient unique values to provide meaningful insights.

Values within the numeric columns are normally distributed and spread out except for the 'number_customer_service_calls,' 'number_vmail_messages,' and 'total_intl_calls' variables, which are skewed to the right, meaning that most of the data values lie to the left of the mean. Variables skewed to the right suggest that most customer calls have lower values for these metrics, with a few having significantly higher values. The dataset shows a balanced distribution of 'area_code', with each code having roughly a third of the entries. Most customers do not have an international plan (no) or a voice mail plan (no). *See the plot below*.

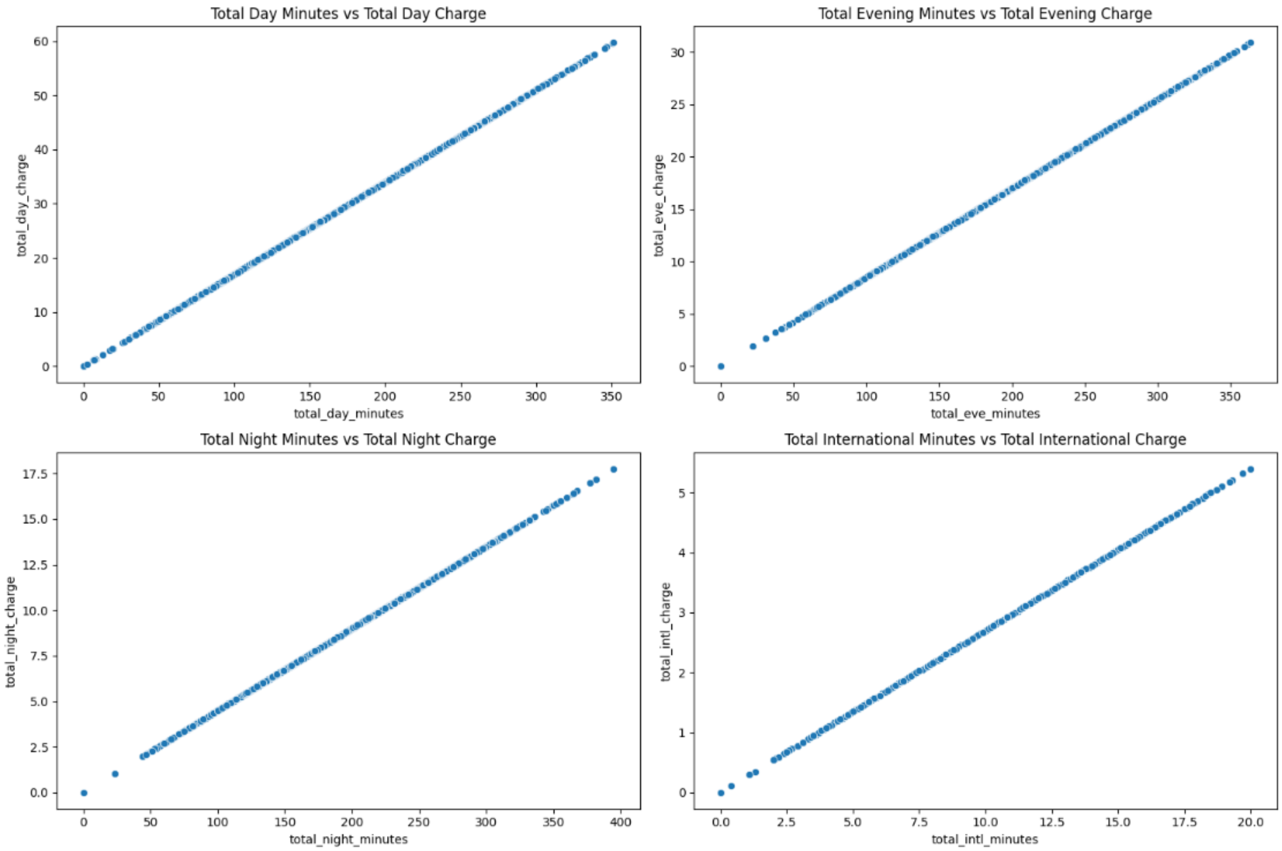




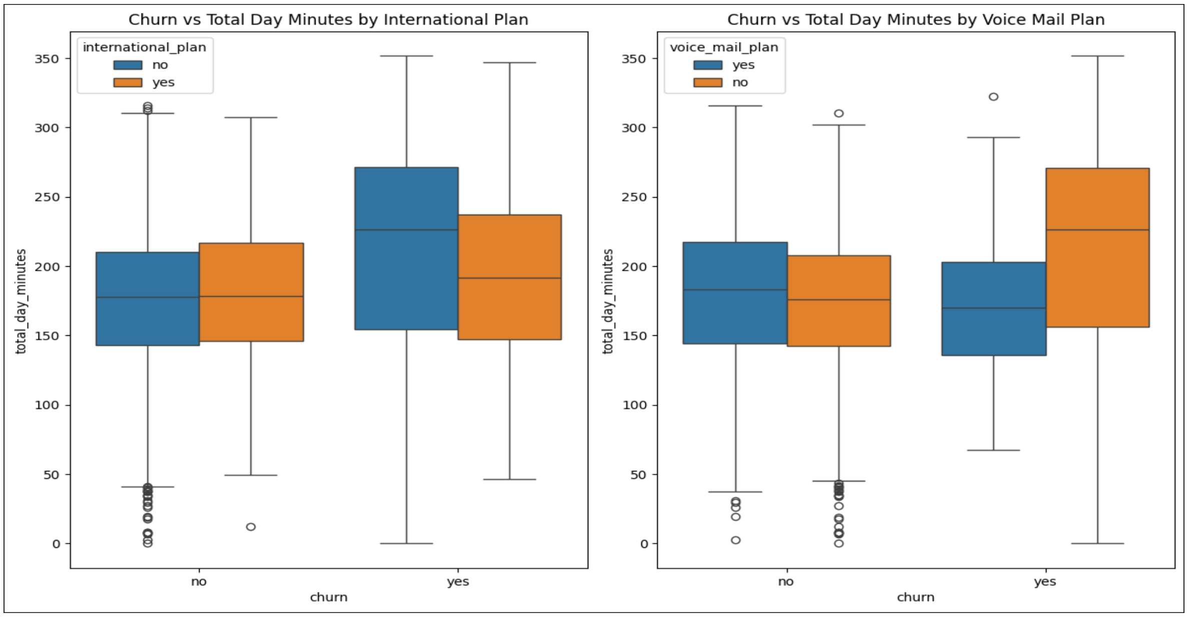


Some variables, like ‘total_day_minutes’ and ‘total_day_charge,’ are closely correlated *(Appendix B)*.

The Relationships between continuous variables revealed a strong linear relationship between minute usage and corresponding charges across different time periods (day, evening, night, international).



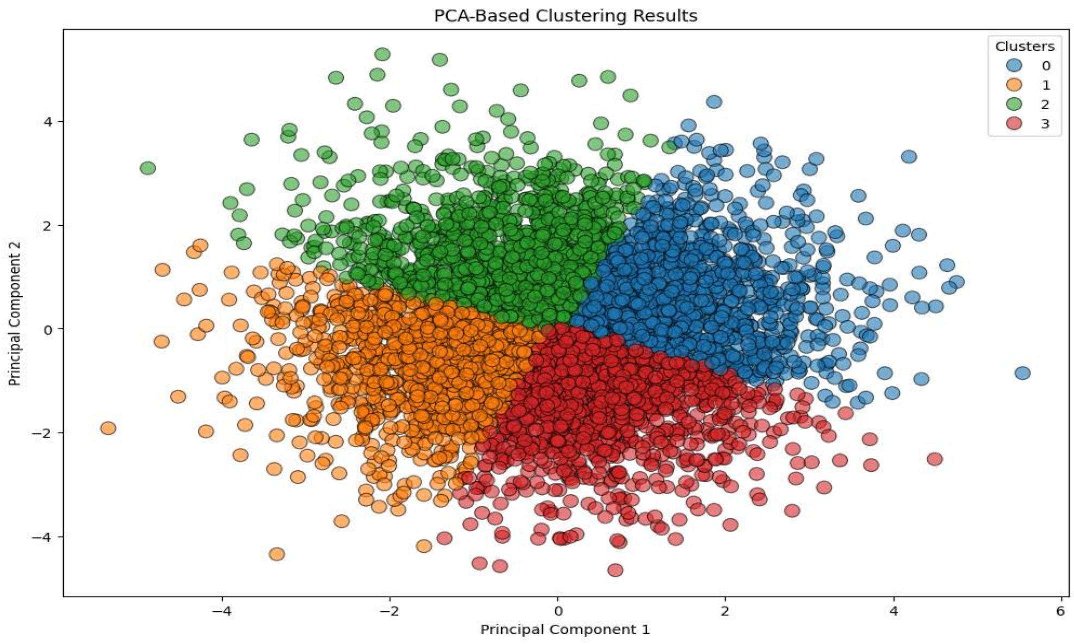
Customers with international plans or voice mail plans show varied usage patterns, which might influence churn event outcomes.



The box plots reveal how the presence of an ‘international_plan’ or ‘voice_mail_plan’ affects the relationship between ‘total_day_minutes’ and ‘churn’, lending credence to the assumption that customers with an international plan or a voicemail plan have different usage patterns compared to those without these plans, especially in relation to the output variable churn. The initial exploration of the dataset reveals essential insights into customer behavior and variable relationships. Because there is a strong correlation between usage and charges and the impact of different plans on probability output. These variables will be crucial in building predictive models for customer churn. The dataset is clean with no missing values, and the categorical variables have sufficient unique values to provide meaningful insights.

Cluster and Principal Component Analysis

EDA was instrumental in identifying trends and relationships within the data. Visualizations such as box plots and correlation matrices helped uncover patterns. For example, EDA showed that higher day minutes and frequent customer service interactions were common among churned customers. Hence, implementing cluster analysis allowed us to compare the outcomes of these unique variables. This analysis provided valuable insights into the customer segments based on continuous variables. Cluster analysis provided additional validation for our findings. Segmenting customers based on key features showed that segments with higher day minutes and customer service interactions had higher churn rates. This consistency between clustering and EDA reinforced the assumptions. However, to verify these assumptions further, Principal Component Analysis PCA was applied to reduce the dataset's dimensionality by generating new uncorrelated variables that sequentially maximize variance (Bera, Pratap, & Verma, 2023). However, some continuous variables did not follow a Gaussian distribution and were transformed to facilitate clustering.



The clusters highlighted the separation and spread of clusters across different dimensions. Each continuous variable grouped by cluster revealed distinct patterns and distributions for different clusters. The number of voicemail messages and customer service calls showed no significant variation across clusters. However, continuous variables like ‘total_day_minutes’ and ‘total_day_charge’ showed varying means across clusters, indicating different customer behaviors, usage patterns, and heavy daytime service usage *(Appendix C)*. The 4 clusters were identified as representing different customer segments with distinct usage patterns. And the transformed variables ‘number_vmail_messages’ and ‘number_customer_service_calls’ helped reduce skewness and improve clustering performance.

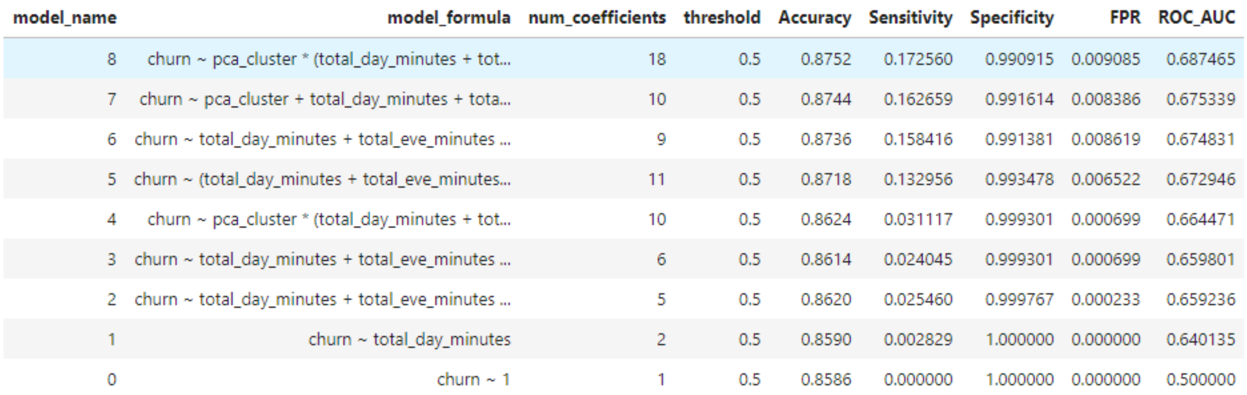
The PCA-based clusters show distinct groupings in the reduced-dimensional space, indicating that PCA effectively captured the variance in the data. The clusters can be interpreted by comparing each cluster's mean values of the original continuous variables. The PCA-based clusters might differ slightly from the original clusters due to the dimensionality reduction, but they still provide valuable insights into the data structure. The most important variables contributing to the PCA-based clusters were identified by examining the loadings of the principal components. These loadings indicate the contribution of each original variable to the principal components. The loadings show how much each original variable contributes to the principal components. Variables with higher absolute values in the loadings are considered more important, thus indicating a stronger association (Faster Capital, n.d.). The First Principal Component (PC1) captures the most variance in the data. Variables with high loadings on PC1 are key in explaining most of the data's variance. The second Principal Component (PC2) captures the second most variance. Variables with high loadings on PC2 add additional insight, often highlighting different aspects of the data.

The variables ‘total_intl_minutes’, ‘total_intl_charge’, ‘total_eve_charge’, and ‘total_eve_minutes’ have high positive loadings on PC1. These variables are crucial in capturing the primary variance in the data, which suggests that international call usage and voicemail messages add another dimension of variance, differentiating customer behavior. The variables ‘total_eve_charge’ , ‘total_eve_minutes', ‘total_day_charge’, ‘total_day_minutes’ have significant loadings on PC2. This indicates that the overall usage of evening minutes and charges across different times is a significant factor in differentiating all the clusters, not just the one captured by PC1 alone. Thus, the combined importance of high loadings on both PC1 and PC2 is especially important. For example, total_eve_minutes and total_eve_minutes are critical for both PC1 and PC2.

Variables like the standardized ‘number_customer_service_calls_log’, which show similar variance in both PCs, contribute moderately to both components, indicating their relevance in explaining variations in customer behavior. Overall, the most important variables identified through PCA are those related to usage and charges across different time periods, international call metrics, and customer service interactions. These variables are key in differentiating customer segments and understanding their behavior, which can inform targeted marketing and service strategies.

Models: Fitting and Interpretation

Model fitting and interpreting involve defining multiple logistic regression models, incorporating combinations of continuous and categorical variables, including interaction terms, and fitting these models using statsmodels. The model coefficients were assessed by evaluating the number of coefficients, their statistical significance, and their significance levels. The performance of each model on the training set was evaluated using metrics such as confusion matrix, accuracy, sensitivity, specificity, false positive rate (FPR), ROC curve, and AUC. The models are compared based on accuracy and ROC AUC to determine which performs best.



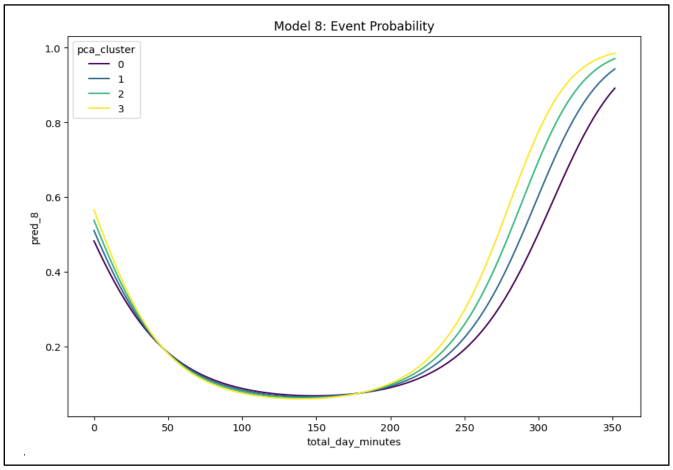
model_name	model_formula	num_coefficients	threshold	Accuracy	Sensitivity	Specificity	FPR	ROC_AUC
8	churn ~ pca_cluster * (total_day_minutes + tot...	18	0.5	0.8752	0.172560	0.990915	0.009085	0.687465
7	churn ~ pca_cluster + total_day_minutes + tota...	10	0.5	0.8744	0.162659	0.991614	0.008386	0.675339
6	churn ~ total_day_minutes + total_eve_minutes ...	9	0.5	0.8736	0.158416	0.991381	0.008619	0.674831
5	churn ~ (total_day_minutes + total_eve_minutes...	11	0.5	0.8718	0.132956	0.993478	0.006522	0.672946
4	churn ~ pca_cluster * (total_day_minutes + tot...	10	0.5	0.8624	0.031117	0.999301	0.000699	0.664471
3	churn ~ total_day_minutes + total_eve_minutes ...	6	0.5	0.8614	0.024045	0.999301	0.000699	0.659801
2	churn ~ total_day_minutes + total_eve_minutes ...	5	0.5	0.8620	0.025460	0.999767	0.000233	0.659236
1	churn ~ total_day_minutes	2	0.5	0.8590	0.002829	1.000000	0.000000	0.640135
0	churn ~ 1	1	0.5	0.8586	0.000000	1.000000	0.000000	0.500000

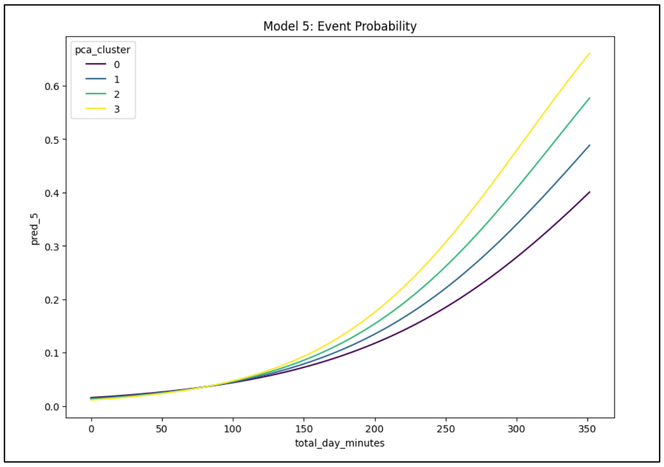
The baseline model, model 0, had low predictive power for churn, but adding total day minutes as a feature improved performance to (ROC AUC of 0.6401). However, combining various types of call minutes further enhanced prediction (ROC AUC around 0.66). The most significant improvement came from introducing interaction terms with PCA clusters and incorporating polynomial and interaction terms, particularly in Model 8, which achieved the highest ROC AUC of 0.6875. These findings highlight the importance of feature engineering and interaction effects for accurately predicting churn. The statistically significant model coefficients were estimated at a p-value less than 0.05, and Model 8 yielded the highest number of 18 coefficients.

Furthermore, the two models with the highest magnitude and classification proportions are Model 8 and Model 5, which correlate with models with the highest statistically significant coefficient. The clusters provide valuable features that enhance churn prediction by capturing distinct customer segments. Evaluating the performance metrics helps understand how well the model predicts churn. Combining clustering and churn prediction enables targeted interventions for high-risk segments, improving customer retention strategies.

Predictions

The prediction results for Models 5 and 8, visualized in the plots, indicate how the event probability (churn) varies with `total_day_minutes`, differentiated by the pca_cluster categories. In both models, the probability of `churn` generally increases as the `total_day_minutes` increases, but the rate of increase varies across different clusters. Specifically, certain clusters exhibit a sharper rise in churn probability with increasing `total_day_minutes`, suggesting that these groups are more sensitive to higher usage. Model 8, which includes interaction terms and polynomial features, shows more nuanced trends with slight fluctuations in the churn probability, indicating a more complex relationship between the features. This suggests that while both models capture the general trend of increasing churn with usage, Model 8 provides a more detailed and potentially more accurate representation of the underlying patterns, especially for different customer segments defined by the PCA clusters.





Insights and Discovery

Churn and Account Length: Customers with shorter account lengths are likelier to churn. Therefore, addressing new customers' needs and expectations can be crucial in reducing churn rates. This pattern suggests that new customers are at higher risk of leaving, possibly due to unmet expectations or initial dissatisfaction.

Impact of Customer Service Calls: High customer service calls are a significant indicator of churn. Thus, improving customer service quality and resolving issues promptly can help retain customers. This pattern is evident across high-risk clusters. Frequent calls to customer service may indicate unresolved issues or dissatisfaction with the service provided.

Usage Patterns and Churn: Therefore, ensuring that heavy users receive reliable service and good value for their usage can help reduce churn.

Plan Types and Churn: The analysis shows that international plans correlate with higher churn rates. Investigating the distribution of international and voice mail plans across different clusters provides insights into customer preferences and potential reasons for churn. This pattern suggests that customers who use international plans might be more sensitive to service issues or pricing, prompting them to seek alternatives.

Distribution of Voicemail Plans: The distribution of voicemail plans across clusters does not significantly impact churn rates. This suggests that voicemail plans might not be a primary factor influencing customer decisions to leave the service.

Geographical Impact: Analysis of churn rates by state (or area) might reveal regional patterns. Certain states or areas could have higher churn rates due to regional service quality issues or competitive market conditions.

Recommendations Based on Findings

1. **Target High-Risk Segments**: Focus retention efforts on customers with high day minutes and frequent customer service calls. Provide personalized offers and proactive support to address their needs.
2. **Improve Customer Service**: Enhance the quality of customer service to resolve issues promptly and reduce dissatisfaction.
3. **Optimize International Plans**: Review and improve international plan offerings to retain customers using these services.

Suggestions for Implementing the Results

1. **Implement in CRM Systems**: Integrate the predictive models into CRM systems to identify at-risk customers and take proactive measures.
2. **Regular Monitoring**: Monitor customer behavior and update models with new data to maintain accuracy.
3. **Customer Feedback**: Use insights to gather feedback from high-risk customers and address their concerns to improve retention.

The following software and modules were used in this project:

- Python (version 3.x)
- Pandas
- NumPy
- Statsmodels
- Scikit-learn
- Seaborn
- Matplotlib
- Anaconda

Acknowledgments

This project was made possible by the contributions of the following individuals and organizations:

Dr. Joseph Yurko
University of Pittsburgh
School of Computing and Information

Dr. Eddie Inyang
Associate Prof.
The College of New Jersey

Anastasia Sosnovskikh,
PPG, Pittsburgh, PA

References

Akhtar, T. (2023, January 12). *Log Transformation and visualizing it using Python*. Retrieved from Medium: https://tariqueakhtar-39220.medium.com/log-transformation-and-visualizing-it-using-python-392cb4bcfc74

Bera, D., Pratap, R., & Verma, B. D. (2023). Dimensionality Reduction for Categorical Data. *IEEE Transactions on Knowledge and Data Engineering, 25*(4), 3658 - 3671.

Faster Capital. (n.d.). *Importance Of Factor Loadings In Data Analysis*. Retrieved from A Faster Capital website: https://fastercapital.com/topics/importance-of-factor-loadings-in-data-analysis.html

Luck, I. (2023, April 4). *What Is Customer Churn? Complete Meaning & Guide*. Retrieved from Customer Guage: https://customergauge.com/customer-churn

Smolic, H. (2023, November 10). *The Ultimate Guide to Understanding and Predicting Customer Churn*. Retrieved from Graphite Note: https://graphite-note.com/the-ultimate-guide-to-understanding-and-predicting-customer-churn/

Appendix A

Dataset information

Number of Rows and Columns:

- The dataset contains **5000 rows** and **20 columns**.

Variable Names and Data Types:

- The dataset includes the following variables with their respective data types:
    - state: object
    - account_length: int64
    - area_code: object
    - international_plan: object
    - voice_mail_plan: object
    - number_vmail_messages: int64
    - total_day_minutes: float64
    - total_day_calls: int64
    - total_day_charge: float64
    - total_eve_minutes: float64
    - total_eve_calls: int64
    - total_eve_charge: float64
    - total_night_minutes: float64
    - total_night_calls: int64
    - total_night_charge: float64
    - total_intl_minutes: float64
    - total_intl_calls: int64

  - total_intl_charge: float64
  - number_customer_service_calls: int64
  - churn: object

Number of Missing Values per Variable:

- There are **no missing values** in the dataset for any variable.

Number of Unique Values per Variable:

- state: 51
- account_length: 212
- area_code: 3
- international_plan: 2
- voice_mail_plan: 2
- number_vmail_messages: 47
- total_day_minutes: 1851
- total_day_calls: 125
- total_day_charge: 1850
- total_eve_minutes: 1671
- total_eve_calls: 120
- total_eve_charge: 1441
- total_night_minutes: 1518
- total_night_calls: 80
- total_night_charge: 990
- total_intl_minutes: 162
- total_intl_calls: 21

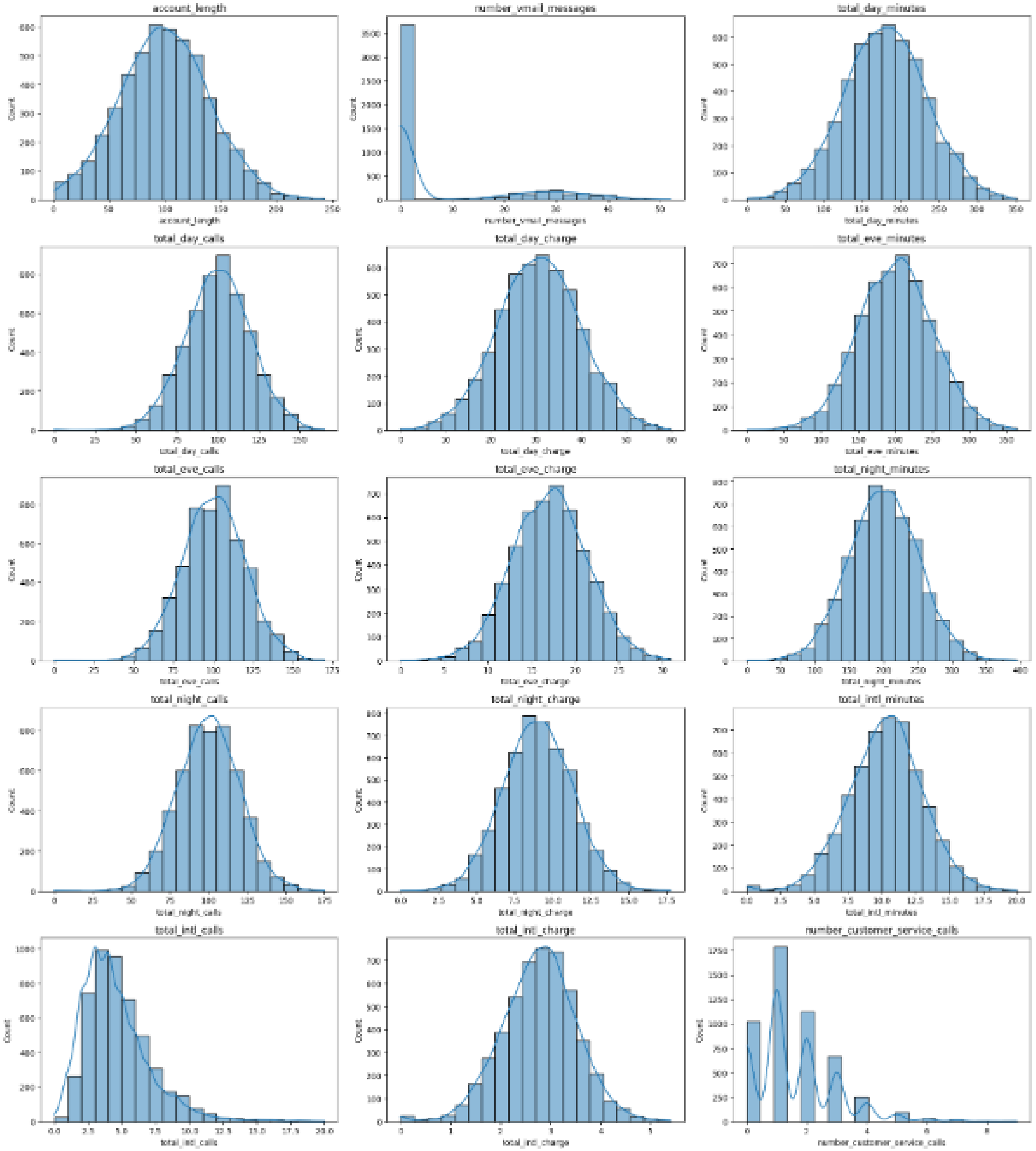
- total_intl_charge: 162
- number_customer_service_calls: 10
- churn: 2

Counts for Categorical Variables:

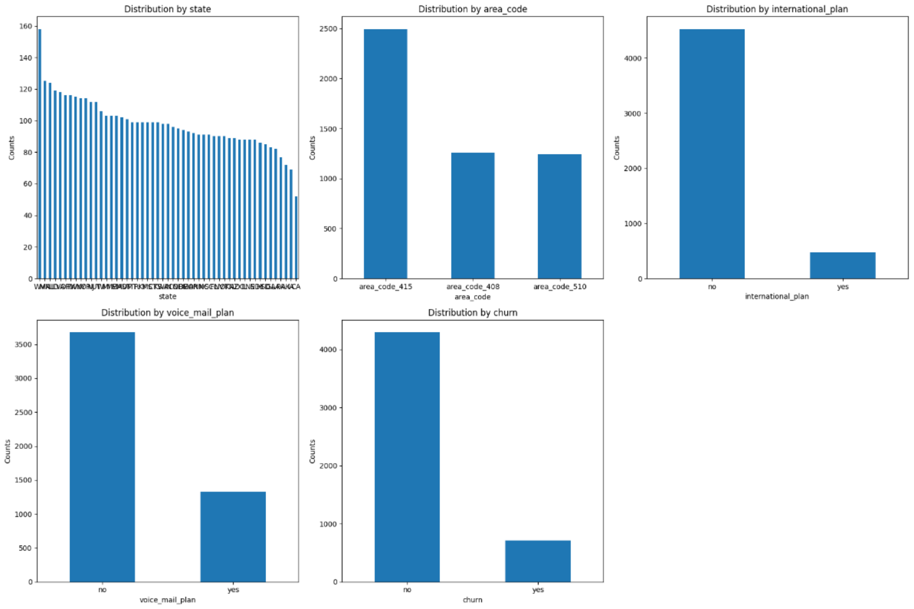
- area_code:
  - 415: 1655
  - 408: 1655
  - 510: 1690
- international_plan:
  - no: 4527
  - yes: 473
- voice_mail_plan:
  - no: 3678
  - yes: 1322
- state: (multiple values, summarized below in visualization)
- churn:
  - false: 4293
  - true: 707

Appendix B

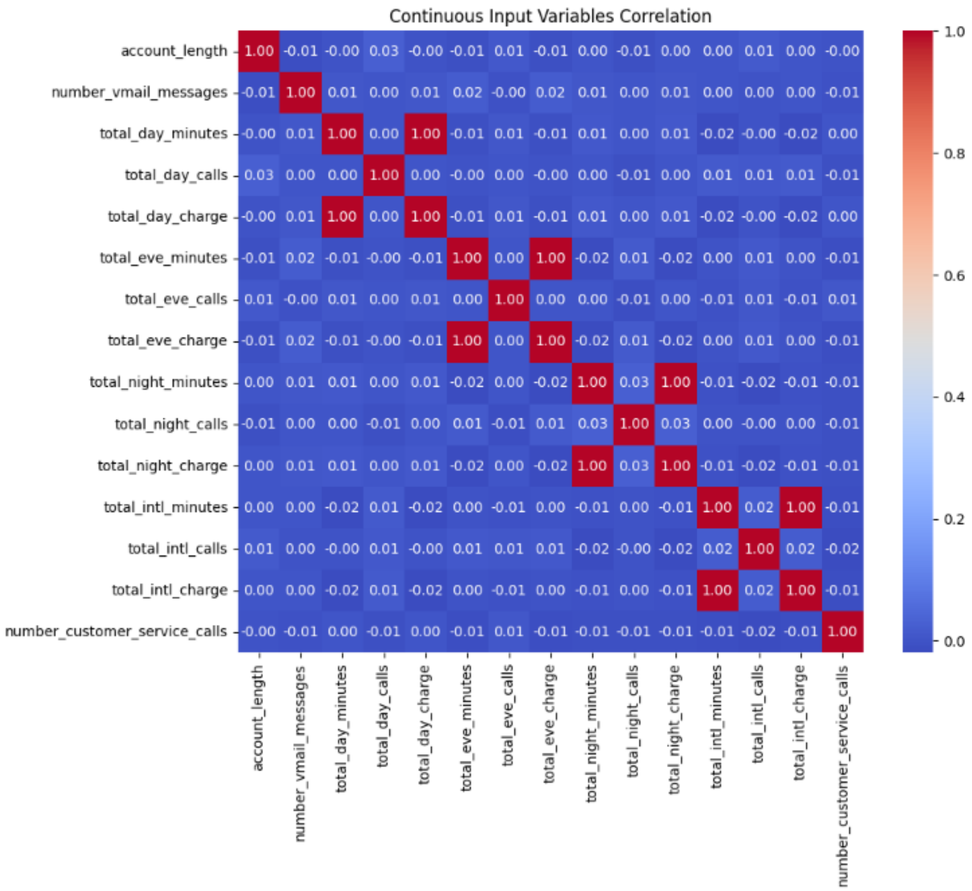
Continuous Variables Distribution



Categorical Variables Distribution



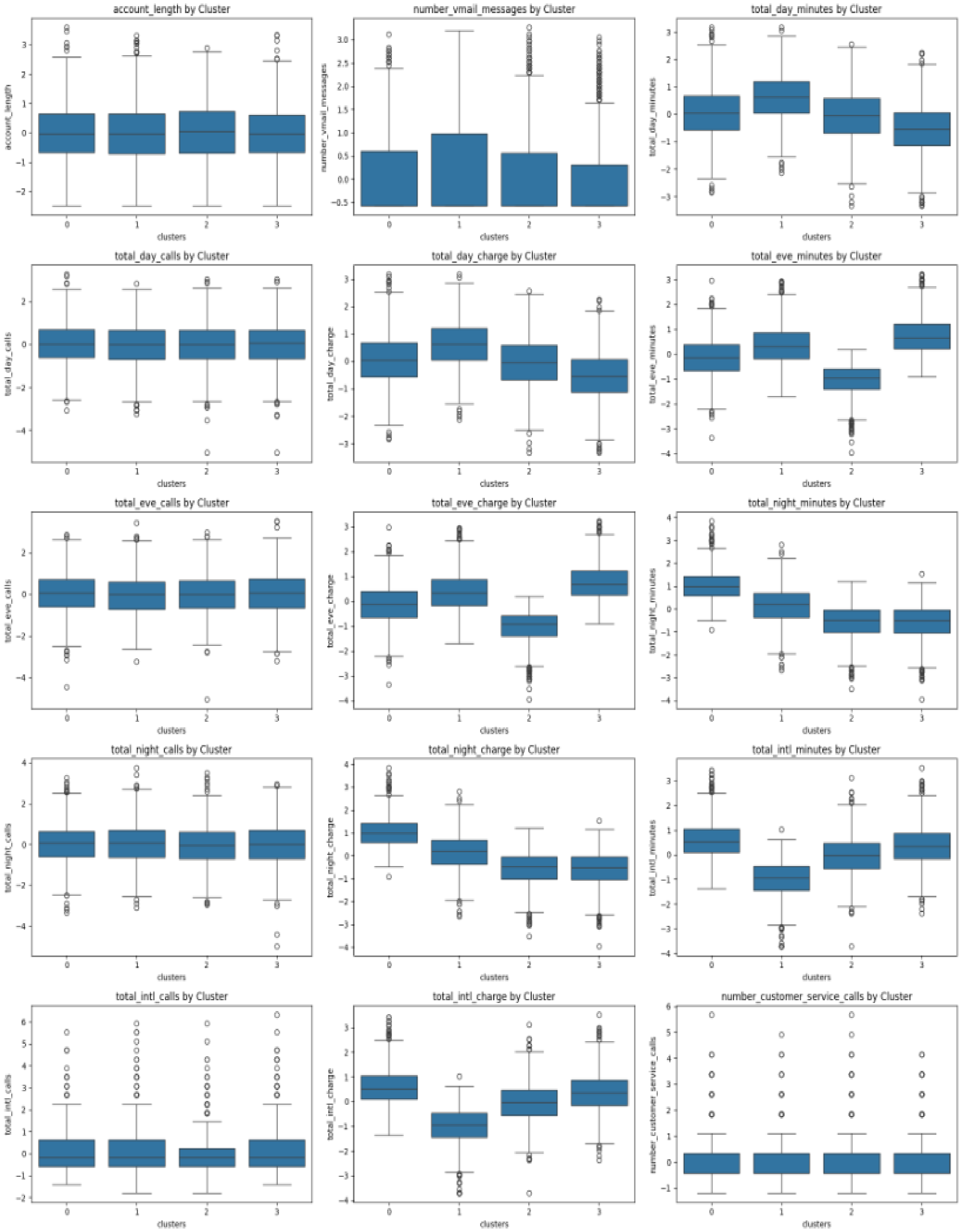
Continuous Variable Correlations



	account_length	number_vmail_messages	total_day_minutes	total_day_calls	total_day_charge	total_eve_minutes	total_eve_calls	total_eve_charge	total_night_minutes	total_night_calls	total_night_charge	total_intl_minutes	total_intl_calls	total_intl_charge	number_customer_service_calls
account_length	1.00	-0.01	-0.00	0.03	-0.00	-0.01	0.01	-0.01	0.00	-0.01	0.00	0.00	0.01	0.00	-0.00
number_vmail_messages	-0.01	1.00	0.01	0.00	0.01	0.02	-0.00	0.02	0.01	0.00	0.01	0.00	0.00	0.00	-0.01
total_day_minutes	-0.00	0.01	1.00	0.00	1.00	-0.01	0.01	-0.01	0.01	0.00	0.01	-0.02	-0.00	-0.02	0.00
total_day_calls	0.03	0.00	0.00	1.00	0.00	-0.00	0.00	-0.00	0.00	-0.01	0.00	0.01	0.01	0.01	-0.01
total_day_charge	-0.00	0.01	1.00	0.00	1.00	-0.01	0.01	-0.01	0.01	0.00	0.01	-0.02	-0.00	-0.02	0.00
total_eve_minutes	-0.01	0.02	-0.01	-0.00	-0.01	1.00	0.00	1.00	-0.02	0.01	-0.02	0.00	0.01	0.00	-0.01
total_eve_calls	0.01	-0.00	0.01	0.00	0.01	0.00	1.00	0.00	0.00	-0.01	0.00	-0.01	0.01	-0.01	0.01
total_eve_charge	-0.01	0.02	-0.01	-0.00	-0.01	1.00	0.00	1.00	-0.02	0.01	-0.02	0.00	0.01	0.00	-0.01
total_night_minutes	0.00	0.01	0.01	0.00	0.01	-0.02	0.00	-0.02	1.00	0.03	1.00	-0.01	-0.02	-0.01	-0.01
total_night_calls	-0.01	0.00	0.00	-0.01	0.00	0.01	-0.01	0.01	0.03	1.00	0.03	0.00	-0.00	0.00	-0.01
total_night_charge	0.00	0.01	0.01	0.00	0.01	-0.02	0.00	-0.02	1.00	0.03	1.00	-0.01	-0.02	-0.01	-0.01
total_intl_minutes	0.00	0.00	-0.02	0.01	-0.02	0.00	-0.01	0.00	-0.01	0.00	-0.01	1.00	0.02	1.00	-0.01
total_intl_calls	0.01	0.00	-0.00	0.01	-0.00	0.01	0.01	0.01	-0.02	-0.00	-0.02	0.02	1.00	0.02	-0.02
total_intl_charge	0.00	0.00	-0.02	0.01	-0.02	0.00	-0.01	0.00	-0.01	0.00	-0.01	1.00	0.02	1.00	-0.01
number_customer_service_calls	-0.00	-0.01	0.00	-0.01	0.00	-0.01	0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	1.00

Appendix C

Principal Component Analysis Features Correlations


account_length by Cluster
account_length
clusters
number_vmail_messages by Cluster
number_vmail_messages
clusters
total_day_minutes by Cluster
total_day_minutes
clusters
total_day_calls by Cluster
total_day_calls
clusters
total_day_charge by Cluster
total_day_charge
clusters
total_eve_minutes by Cluster
total_eve_minutes
clusters
total_eve_calls by Cluster
total_eve_calls
clusters
total_eve_charge by Cluster
total_eve_charge
clusters
total_night_minutes by Cluster
total_night_minutes
clusters
total_night_calls by Cluster
total_night_calls
clusters
total_night_charge by Cluster
total_night_charge
clusters
total_intl_minutes by Cluster
total_intl_minutes
clusters
total_intl_calls by Cluster
total_intl_calls
clusters
total_intl_charge by Cluster
total_intl_charge
clusters
number_customer_service_calls by Cluster
number_customer_service_calls
clusters