Customer Segmentation and Churn Prediction in Telecommunications Companies

CS5228 Knowledge Discovery and Data Mining Course Mini Project

Contributors:
- Dou Maokang (A0210799W) - EDA and Preprocessing
- Euan Rodger (A0304973Y) - EDA and Preprocessing
- Muhammad Riandi Ramadhan (A0314534L) - Unsupervised
- Xiao Yicong (A0304728A) - Supervised

NUS National University of Singapore

DATASET & OBJECTIVE

Dataset

The study is based on customer records drawn from a U.S. telecom provider (80 % training, 20 % test). Each row corresponds to a single account and includes 19 explanatory variables plus a binary churn label. For clarity, the features can be grouped as follows:

- Usage metrics: Total minutes, number of calls, and total charge are reported separately for the day, evening, night, and international time-bands. These variables reveal both overall traffic volume and time-of-day preferences.
- Plan attributes: Binary flags indicate whether the customer subscribes to an international calling plan or a voice-mail plan.
- Tenure and geography: Account length, area code, and state capture longevity and location-based factors.

	Size
Train	2666
Test	667

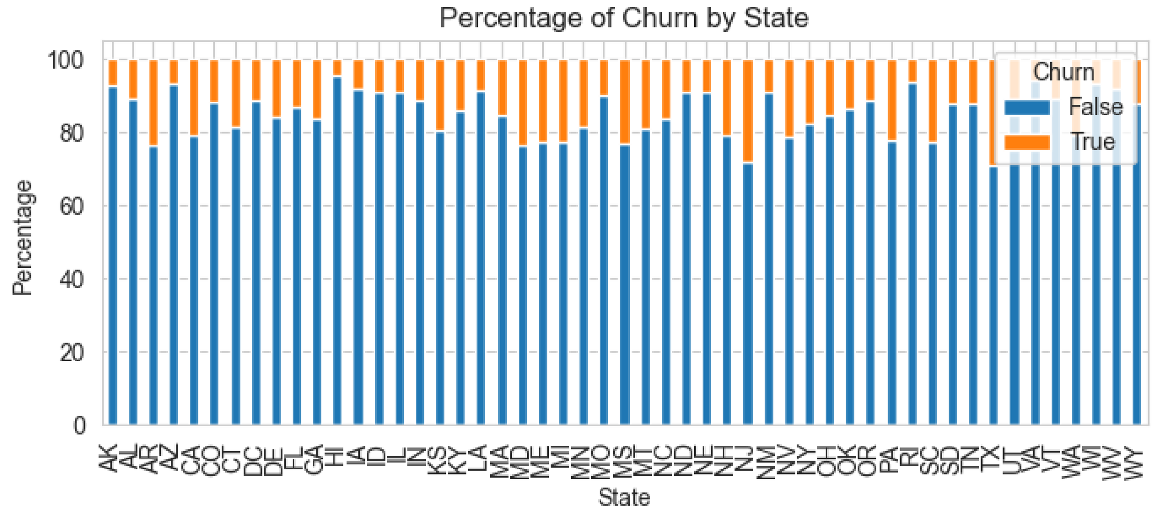
Churn	#count in train set
No	2278
Yes	388

Objective

Our primary objectives are to segment customers to identify distinct groups, and to predict customer churn based on existing data.

EDA

Categorical Features



How Categorical Features Contribute to Churn?

The churn rate is visualized for categorical features such as State, Area Code, International Plan and Voice-Mail Plan.

State

- The churn rate fluctuate within 20% across all states.
- The variation appears noisy and is probably driven by small sample sizes in many states.

Area Code

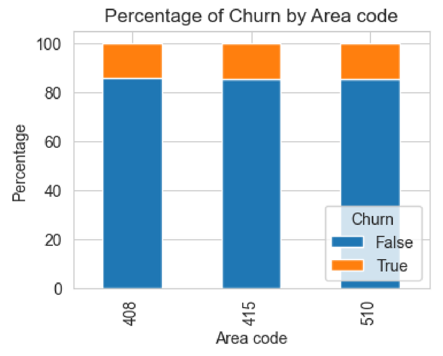
- The three area codes exhibit very similar churn fractions.
- Because their churn rates are essentially indistinguishable, Area Code carries little standalone predictive value.

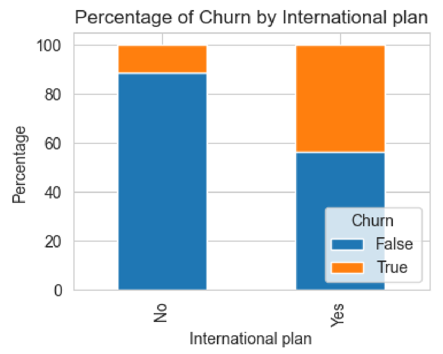
International Plan

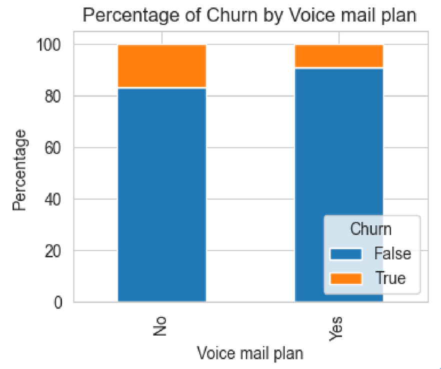
- Customers who subscribe to an international plan churn at roughly 42 percent, compared with just 12 percent for those without the plan.
- This three-to-four-fold lift makes the International Plan flag the single most decisive categorical indicator of attrition.

Voice-Mail Plan

- Holding a voice-mail plan is associated with a drop in churn, from about 15 percent to 9 percent.
- The plan appears to be protective, possibly because fewer missed calls translate into fewer service issues and complaints.







EDA & PREPROCESSING

Numerical Features

How Numerical Features Contribute to Churn?

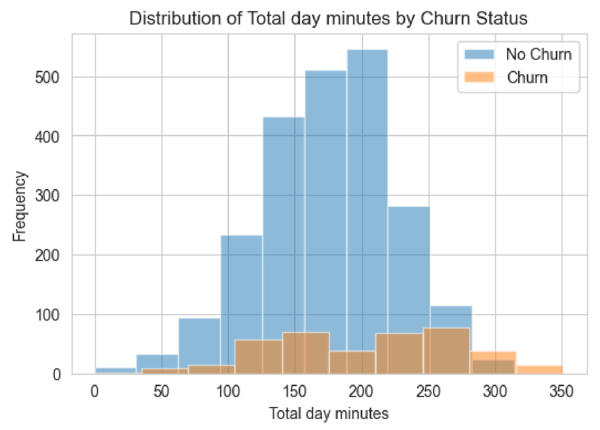
In the section below, we present the distributions of several numerical features for churners and non-churners using overlay histograms. We observe that the distributions are broadly similar between the two groups for most variables, whereas certain features—such as Total Day Minutes and Number of Customer Service Calls—show clearly distinct patterns.

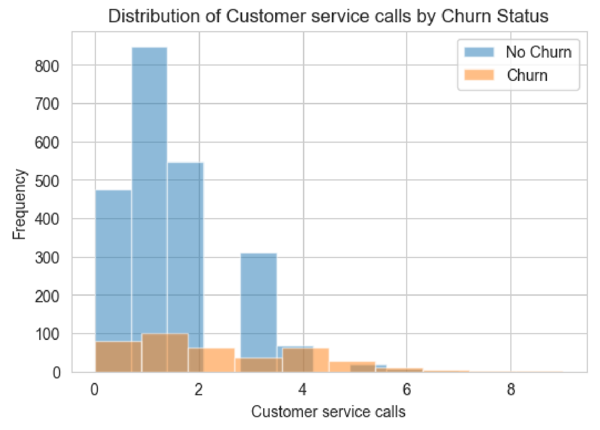
Total day minute – "heavier use, higher churn"

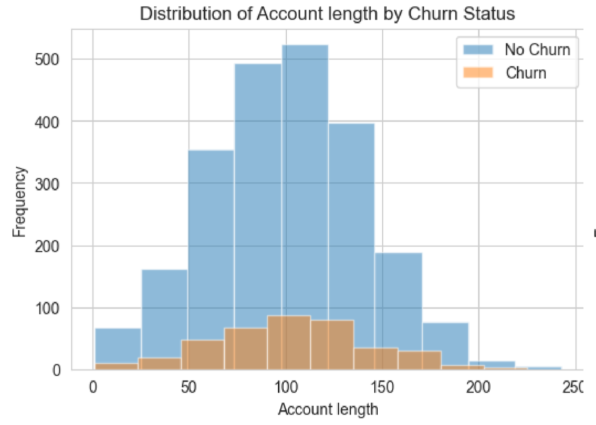
- Usage by churners is skewed towards the right.
- By 250 minutes the churn share more than doubles relative to light users.
- By 350 minutes it's almost all churners.

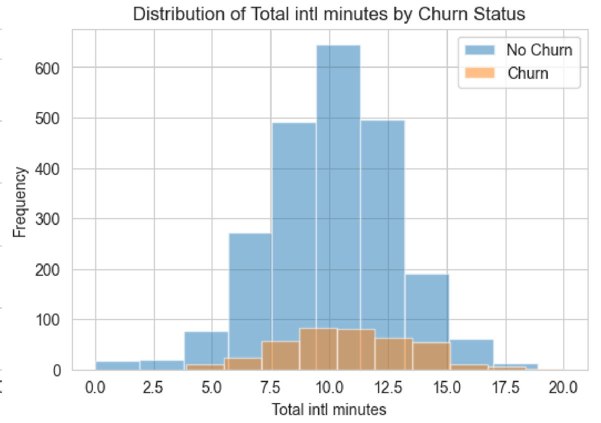
Customer-service calls – "a sharp dissatisfaction threshold"

- Up to three calls, churners are a minority.
- At four or more calls the pattern flips, with churners outnumbering stayers despite being a small fraction of the overall base.









Preprocessing

Encoding

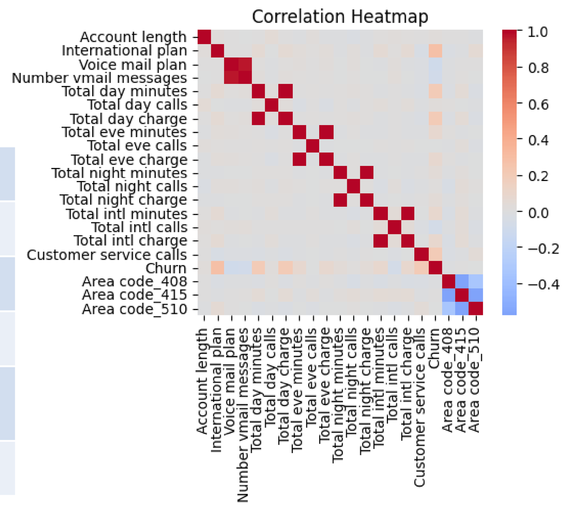
- Nominal features (Area code, State) encoded using one-hot, creating a new binary feature for each state and area code.
- Binary Yes/No features encoded as 1 or 0.
- Numerical features normalized via Min-Max scaler (fit between [0, 1]), based on the values in the training data.

Feature Selection

- Correlation analysis was performed to identify the most significant features and any potential collinearity issues.
- The 'Total X minutes' and 'Number vmail messages' features were removed due to collinearity.

Highest correlation feature pairs:

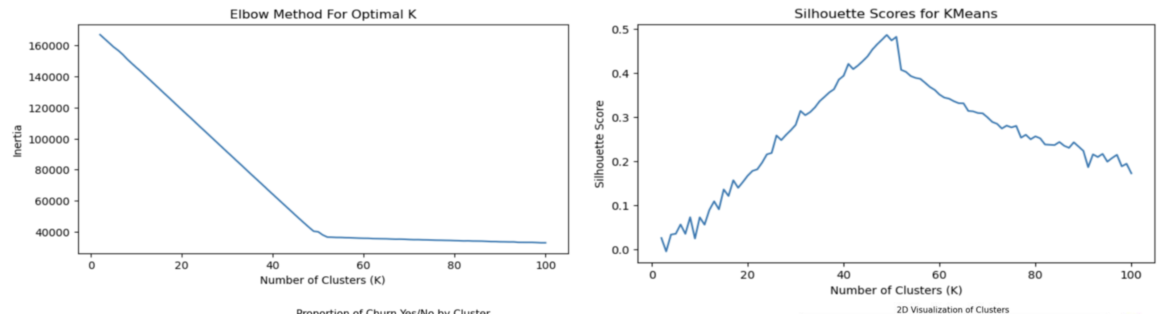
Total day charge	Total day minutes	1.000
Total eve charge	Total eve minutes	1.000
Total night charge	Total night minutes	0.999
Total intl charge	Total intl minutes	0.999
Number vmail messages	Voice mail plan	0.957
Next highest correlated pair		0.277

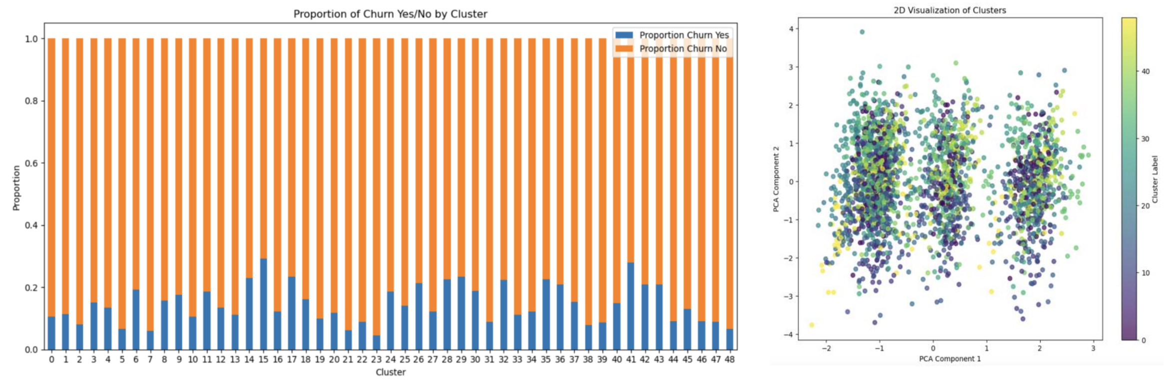


State columns omitted for having no standout features

UNSUPERVISED LEARNING

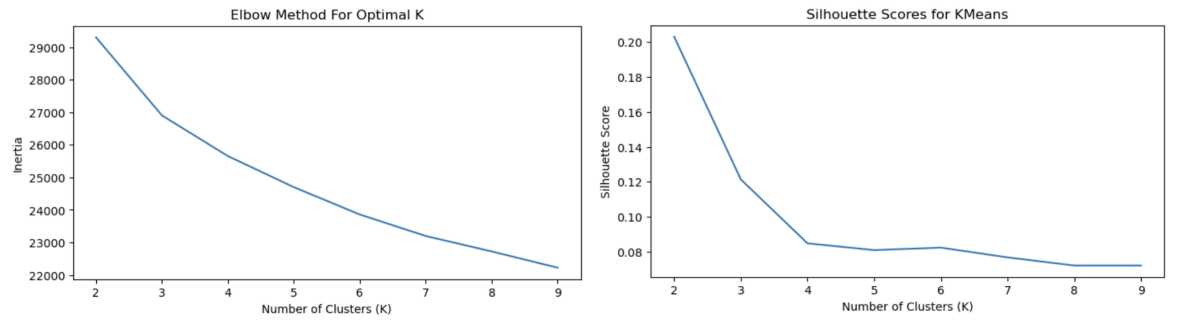
KMEANS

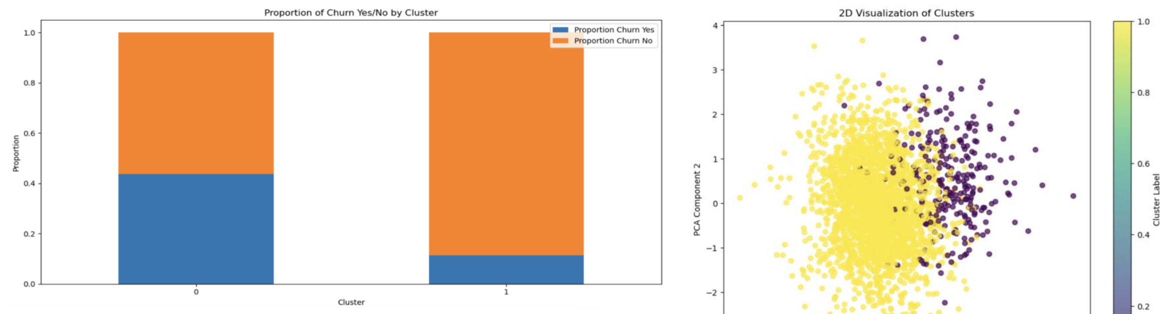




- The optimal number of cluster is **49** (inertia = 40283.83, Silhouette score = 0.486). Each cluster has 24-128 data in varies.
- After grouping each cluster data, we found that encoded **"state"** columns have 0 standard deviation (std) which shows **no variability**. There is an exception for 1 cluster which has >0 std in 2 states.
- It is suggested that the cluster is formed based on the state column value. Therefore, we try to **remove location related columns** (state and area code) for comparison.

After "state" and "area code" are dropped



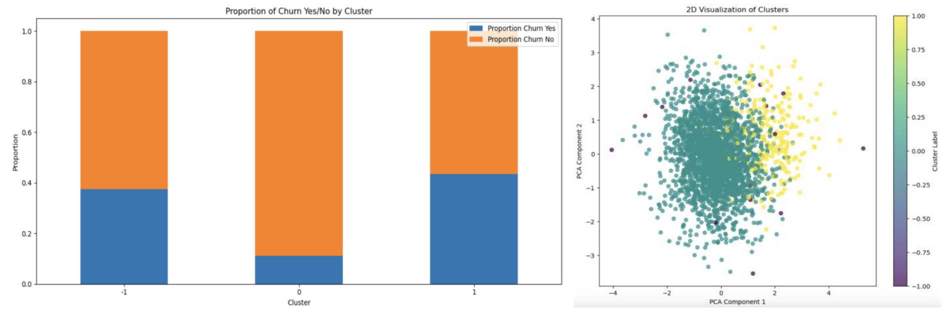


Cluster	Total Members	Churn Yes Count	Proportion Churn Yes
0	270	118	0.437037
1	2396	270	0.112688

- The optimal number of cluster is **2** (inertia = 29303.16, Silhouette score = 0.203).
- After grouping each cluster data, we found that encoded **"international plan"** column has 0 standard deviation (std) which shows **no variability**. It is suggested that the cluster is formed based on this.

DBSCAN

To tune hyperparameters, we utilized grid search with **eps range** of **3.0 - 5.0** and **min samples range** of **10-20**. We also dropped "state" and "area code" columns since for the clustering.



eps = **~3.421**
min_samples = **10**
silhouette = **0.198**

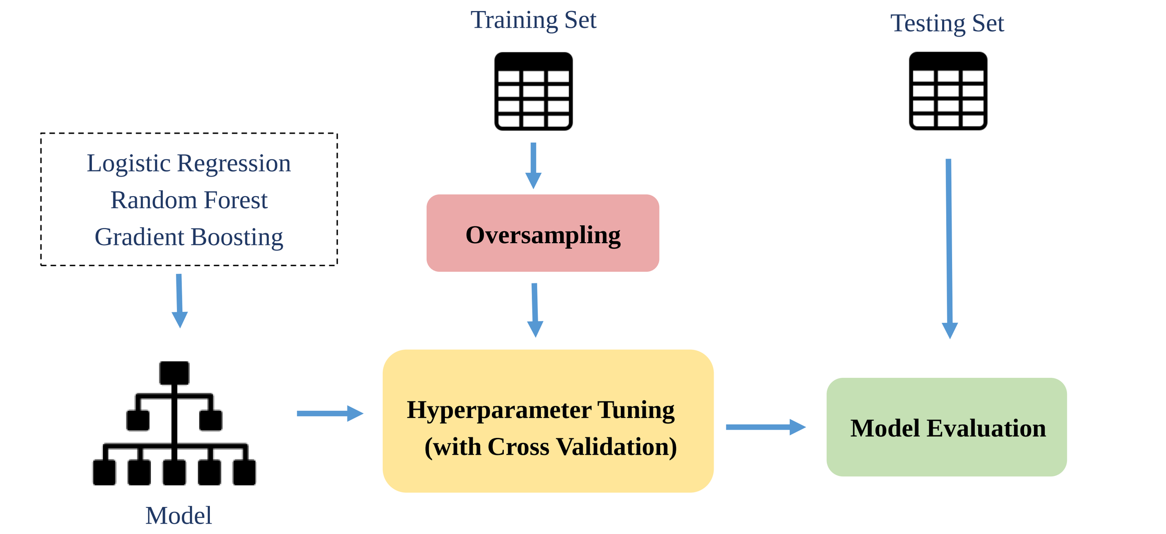
It identifies that there are **2 clusters** with **16 outliers**.

Cluster	Total Members	Churn Yes Count	Proportion Churn Yes
1	262	114	0.435115
-1	16	6	0.375000
0	2388	268	0.112228

Same as KMeans, **"international plan"** column has 0 standard deviation (std) which shows **no variability**. It is suggested that the cluster is formed based on this. Each cluster's **"churn"** proportions are similar with KMeans too.

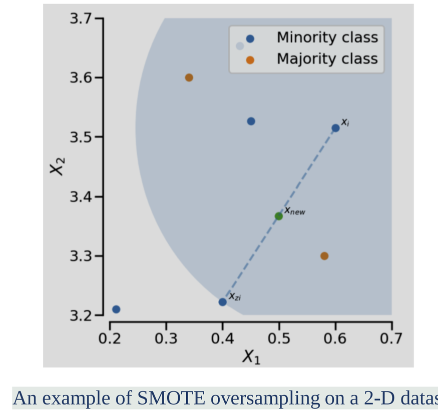
SUPERVISED ALGORITHM TRAINING

Overall Procedure



Oversampling

- The dataset is imbalanced, as only **14.6%** of customers in the train set chose to chur.
- We performed **SMOTE** oversampling on the **training set** to generate synthetic "customers" that churn.



An example of SMOTE oversampling on a 2-D dataset

Churn	#count
No	2278
Yes	388

Churn	#count
No	2278
Yes	2278

- This oversampling stage improved the inference performance of all the models we trained. A result comparison is shown in later section.

Hyperparameter Tuning

- Logistic Regression

Parameter	Motivation	Value Range
Scale of penalty terms	Regularization Behavior	[0.001, 0.01, 0.1, 1, 10, 100]
Penalty Loss		L1, L2, Elastic-Net

- Random Forest

Parameter	Motivation	Value Range
Number of estimator	Balance complexity vs. predictive power	[50, 100, 200]
Maximum depth of trees	Control overfitting	[No Restriction, 10, 20, 30]

- Gradient Boosting

Parameter	Motivation	Value Range
Number of estimator	Balance complexity vs. predictive power	[50, 100, 200]
Learning rate	Control overfitting	[0.01, 0.05, 0.1, 0.2]
Sampling ratio for rows in each iteration		[0.5, 0.6, 0.7, 0.8, 0.9, 1.0]
Sampling ratio for cols in each iteration		
Maximum depth of trees		[6, 10, 20, 30]

Grid Search with 5-Fold Cross Validation

Select the hyperparameter combination that gives the best average **F1 Score** across the 5 validation sets

SUPERVISED ALGORITHM RESULT ANALYSIS

Ablation Study

- **Oversampling**: The oversampling step enhanced the **recall** on test set for logistic regression and random forest

	Recall		Precision		F1	
	Regular	Trained on oversampled data	Regular	Trained on oversampled data	Regular	Trained on oversampled data
Logistic Regression	0.2526	**0.6** (+0.34)	0.5106	0.3958	0.3380	0.477
Random Forest	0.6	**0.7263** (+0.1263)	0.9828	0.7582	0.7451	0.7419
Gradient Boosting	**0.7895**	0.7684	0.8929	0.8202	0.8380	0.7935

- **Removing the Location Features**: Removing the "State_xx" and "Area code_xx" features from both the training and testing set further enhanced result (trained on oversampled data)

	Recall		Precision		F1	
	With location	W/o location	With location	W/o location	With location	W/o location
Logistic Regression	0.6	**0.7579** (+0.16)	**0.3958**	0.3636	0.477	**0.4915** (+0.01)
Random Forest	0.7263	**0.8** (+0.07)	0.7582	**0.8172** (+0.06)	0.7419	**0.8085** (+0.06)
Gradient Boosting	**0.7684**	0.7579	0.8202	**0.8571** (+0.0)	0.7935	**0.8045**

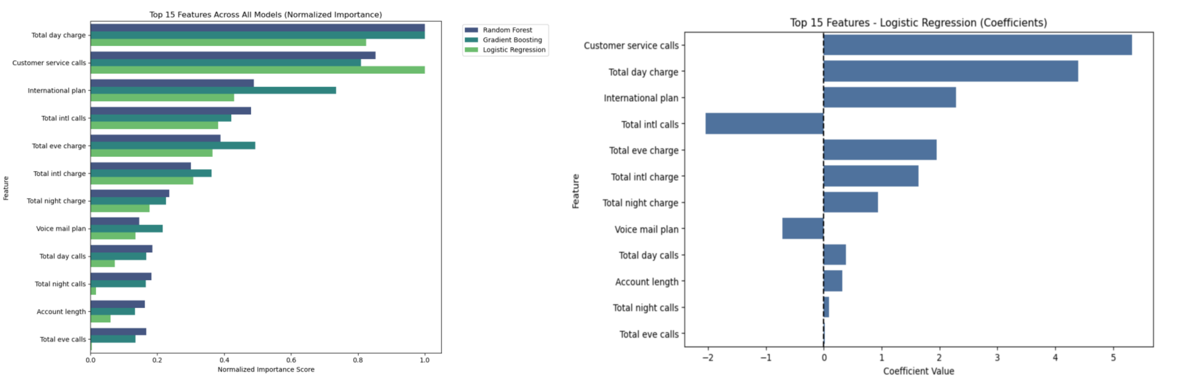
Feature Importance

We define the following importance scores for each feature:

- **Logistic Regression Feature Score**:
  We directly use the coefficient as indicator of feature importance
- **Random Forest/Gradient Boosting Feature Score**:
  Average reduction of impurity across all trees

$$I_j = \frac{1}{N_{\text{trees}}}\sum_{T=1}^{N_{\text{trees}}}\left(\sum_{n\in T\,\text{split on}\,j}\Delta\text{Impurity}(n)\right)$$

For each feature, we apply normalization to its scores obtained from the two measures, and use the coefficients of logistic regression to indicate whether the feature positively or negatively contribute to churn rate.



Insight and Recommendation

1. Be Responsive to Service Calls

This is a decisive feature in all supervised training algorithms, since increasing number of service calls possibly indicate larger resentment from customers. The company can try to investigate common reason for service calls to find out the root cause for customer dissatisfaction.

2. Inspect the Pricing of International Plan

Both our unsupervised and supervised learning results indicate a strong correlation between subscription to international plan and churning. However, the number of international calls is negatively contributing to churn rate. This seemingly conflicting result may indicate a mismatch between the utility of having ad-hoc international calls v.s. having a long-term international plan. For instance, customers may find it more cost-affordable to make ad-hoc international calls than purchasing a plan. This suggests a potential need for more market research on the pricing strategy of international plans.

3. Consider Price Discount for Active Users

Various phone usage related features (e.g. "total day charge", "account length") are positively correlated with churn rate. This may be due to more affordable plan offered by competitors for more active customers. The company can do more market research on the pricing of telecom service and offer loyalty plan/discount to frequent users.