# Data Preparation & Exploratory Data Analysis (EDA)

## Data Dictionary

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Explanation** | **Variable Type** | **Data Type** |
| **State** | Customer’s state | Categorical (Nominal) | object |
| **Account Length** | Integer number showing the duration of activity for customer account | Quantitative (Continuous) | int64 |
| **Area Code** | Area code of customer | Categorical (Nominal) | int64 |
| **Phone Number** | Phone number of customer | Categorical (Nominal) | object |
| **Inter Plan** | Binary indicator showing whether the customer has international calling plan | Categorical/ Binary (yes, no) | object |
| **VoiceMail Plan** | Indicator of voice mail plan | Categorical/ Binary (yes, no) | object |
| **No of Vmail Mesgs** | The number of voicemail messages | Quantitative (Discrete) | int64 |
| **Total Day Min** | The number of minutes the customer used the service during day time | Quantitative (Continuous) | float64 |
| **Total Day calls** | Discrete attribute indicating the total number of calls during day time | Quantitative (Discrete) | int64 |
| **Total Day Charge** | Charges for using the service during day time | Quantitative (Continuous) | float64 |
| **Total Evening Min** | The number of minutes the customer used the service during evening time | Quantitative (Continuous) | float64 |
| **Total Evening Calls** | The number of calls during evening time | Quantitative (Discrete) | int64 |
| **Total Evening Charge** | Charges for using the service during evening time | Quantitative (Continuous) | float64 |
| **Total Night Minutes** | Number of minutes the customer used the service during night time | Quantitative (Continuous) | float64 |
| **Total Night Calls** | The number of calls during night time | Quantitative (Discrete) | int64 |
| **Total Night Charge** | Charges for using the service during night time | Quantitative (Continuous) | float64 |
| **Total Int Min** | Number of minutes the customer used the service to make international calls | Quantitative (Continuous) | float64 |
| **Total Int Calls** | The number of international calls | Quantitative (Discrete) | int64 |
| **Total Int Charge** | Charges for international calls | Quantitative (Continuous) | float64 |
| **No of Calls Customer Service** | The number of calls to customer support service | Quantitative (Discrete) | int64 |
| **Churn** | Class attribute with binary values (True for churn and False for not churn) | Categorical/ Binary (TRUE, FALSE) | object |

Categorical columns are - State, Area Code, Phone Number, Inter Plan, VoiceMail Plan, and Churn.

Rest : are quantitative/numerical

Use encoding technique to transform those categorical columns into numerical values, before building predictive models by machine learning.

## Missing Values & Duplications

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 3333 entries, 0 to 3332

Data columns (total 21 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 State 3333 non-null object

1 Account Length 3333 non-null float64

2 Area Code 3333 non-null object

3 Phone Number 3333 non-null object

4 Inter Plan 3333 non-null object

5 VoiceMail Plan 3333 non-null object

6 No of Vmail Mesgs 3333 non-null float64

7 Total Day Min 3333 non-null float64

8 Total Day calls 3333 non-null float64

9 Total Day Charge 3333 non-null float64

10 Total Evening Min 3333 non-null float64

11 Total Evening Calls 3333 non-null float64

12 Total Evening Charge 3333 non-null float64

13 Total Night Minutes 3333 non-null float64

14 Total Night Calls 3333 non-null float64

15 Total Night Charge 3333 non-null float64

16 Total Int Min 3333 non-null float64

17 Total Int Calls 3333 non-null float64

18 Total Int Charge 3333 non-null float64

19 No of Calls Customer Service 3333 non-null float64

20 Churn 3333 non-null object

dtypes: float64(15), object(6)

memory usage: 546.9+ KB

Counting the number of missing values for each column by df.isna().sum()

State 0

Account Length 0

Area Code 0

Phone Number 0

Inter Plan 0

VoiceMail Plan 0

No of Vmail Mesgs 0

Total Day Min 0

Total Day calls 0

Total Day Charge 0

Total Evening Min 0

Total Evening Calls 0

Total Evening Charge 0

Total Night Minutes 0

Total Night Calls 0

Total Night Charge 0

Total Int Min 0

Total Int Calls 0

Total Int Charge 0

No of Calls Customer Service 0

Churn 0

dtype: int64

Counting the number of duplicated rows for the dataset by df.duplicated().value\_counts()

False 3333

dtype: int64

Found no missing values and duplicates in the dataset.

### Statistical Summary for numerical columns (max, min, mean, standard deviation)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **index** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **Account Length** | 3333.0 | 101.0648065 | 39.82210593 | 1.00 | 74.00 | 101.00 | 127.00 | 243.00 |
| **No of Vmail Mesgs** | 3333.0 | 8.099009901 | 13.68836537 | 0.00 | 0.00 | 0.00 | 20.00 | 51.00 |
| **Total Day Min** | 3333.0 | 179.7750975 | 54.4673892 | 0.00 | 143.70 | 179.40 | 216.40 | 350.80 |
| **Total Day calls** | 3333.0 | 100.4356436 | 20.06908421 | 0.00 | 87.00 | 101.00 | 114.00 | 165.00 |
| **Total Day Charge** | 3333.0 | 30.56230723 | 9.259434554 | 0.00 | 24.43 | 30.50 | 36.79 | 59.64 |
| **Total Evening Min** | 3333.0 | 200.980348 | 50.71384443 | 0.00 | 166.60 | 201.40 | 235.30 | 363.70 |
| **Total Evening Calls** | 3333.0 | 100.1143114 | 19.92262529 | 0.00 | 87.00 | 100.00 | 114.00 | 170.00 |
| **Total Evening Charge** | 3333.0 | 17.08354035 | 4.310667643 | 0.00 | 14.16 | 17.12 | 20.00 | 30.91 |
| **Total Night Minutes** | 3333.0 | 200.8720372 | 50.57384701 | 23.20 | 167.00 | 201.20 | 235.30 | 395.00 |
| **Total Night Calls** | 3333.0 | 100.1077108 | 19.56860935 | 33.00 | 87.00 | 100.00 | 113.00 | 175.00 |
| **Total Night Charge** | 3333.0 | 9.039324932 | 2.275872838 | 1.04 | 7.52 | 9.05 | 10.59 | 17.77 |
| **Total Int Min** | 3333.0 | 10.23729373 | 2.791839548 | 0.00 | 8.50 | 10.30 | 12.10 | 20.00 |
| **Total Int Calls** | 3333.0 | 4.479447945 | 2.461214271 | 0.00 | 3.00 | 4.00 | 6.00 | 20.00 |
| **Total Int Charge** | 3333.0 | 2.764581458 | 0.753772613 | 0.00 | 2.30 | 2.78 | 3.27 | 5.40 |
| **No of Calls Customer Service** | 3333.0 | 1.562856286 | 1.315491045 | 0.00 | 1.00 | 1.00 | 2.00 | 9.00 |

11 out of 15 of the numeric attributes have a minimum value of 0

## Outliers Detection for numerical columns

Show the outliers of the attributes using boxplot:

Chart

Description automatically generatedChart

Description automatically generated

Chart, box and whisker chart

Description automatically generatedChart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart

Description automatically generated

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Chart

Description automatically generatedChart

Description automatically generatedChart

Description automatically generated

Chart

Description automatically generated

outliers across columns is not significant, so as per our last discussion we will not drop the outliers

### . Distribution Visualization of numerical columns & Impacts on Class Attribute (Bonus)

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

most of the numeric variables have a pretty much symmetrical, unimodal distribution as a bell curve

exceptions: No of Vmail Mesgs, Total Int Calls, No of Calls Customer Service

**The “No of Vmail Mesgs” attribute:**

Chart, histogram

Description automatically generated

Graphical user interface, application

Description automatically generatedA picture containing text, shoji

Description automatically generated

“No of Vmail Mesgs”does not impact much on the class attribute

**The “Total Int Calls” & “No of Calls Customer Service” attributes:** skewed to the right

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated Chart, histogram

Description automatically generated

there are proportionate differences

class\_FALSE

normal distribution in the “Total Int Calls”

bimodal distributed in the “No of Calls Customer Service”

class\_TRUE

left side weights are relatively small

keeping findings as a reference point: just for now: will discuss with Eric and Raymond

## Correlation Heatmap (Bonus)

**categorical columns excluded (**haven’t transformed them into numerical values)

Chart, histogram

Description automatically generated

correlation heatmap shows perfect correlation

***Total Day Charge – Total Day Min,***

***Total Evening Charge – Total Evening Min,***

***Total Night Charge – Total Night Minutes,***

***Total Intl Charge - Total Int Min***.

## Class Imbalance in “Churn” – churned (TRUE) or not churned (FALSE) (Bonus)

Chart

Description automatically generated with medium confidence

Percentage of churned customer: 0.14491449144914492

Percentage of not-churned customer: 0.8550855085508551

**Calculating the current churn-rate:**

Churn Rate: 0.14491449144914492

Churned : 15 %

No- Churned: 85%

## Categorical to Numeric Encoding for categorical columns (Bonus)

“Phone Number”- Drop – No specific significance in analysis

One hot Encoding technique

categorical attributes**: "State", "Area Code", "Inter Plan", "VoiceMail Plan".**

Still working on this

# Predictive Modeling (Classification)

## Data Splitting Strategy

the 70% training and 30% testing spit method

## . Classification using Decision Tree (Supervised Learning)

**The Decision Tree baseline model:**

A picture containing table

Description automatically generated

The Baseline Classification Tree Model accuracy score **is**: 0.9270

precision recall f1-score support

FALSE 0.96 0.96 0.96 850

TRUE 0.76 0.75 0.76 150

accuracy 0.93 1000

macro avg 0.86 0.86 0.86 1000

weighted avg 0.93 0.93 0.93 1000

The confusion matrix for the baseline decision tree model:

Chart, treemap chart

Description automatically generated

The overall accuracy of this baseline decision tree is 92.7% ~ 93% which is great

Precision and recall are not much different, respectively at 76% and 75%.

**The Decision Tree model with selected features:**

# Highly correlated features (8 columns) to be excluded

highly\_cor\_feature\_names=["Total Day Min", "Total Day Charge", "Total Evening Min", "Total Evening Charge", "Total Night Minutes", "Total Night Charge", "Total Int Min", "Total Int Charge"]

Total Day calls 4000 non-null float64

Total Evening Calls 4000 non-null float64

Total Int Calls 4000 non-null float64

Total Night Calls 4000 non-null float64

VoiceMail Plan\_no 4000 non-null uint8

VoiceMail Plan\_yes 4000 non-null uint8

Attributes such as***: Inter Plan, Total Int Calls, No of Calls Customer Service, Total Day calls, No of Calls Customer Service, VoiceMail Plan, Area Code\_A510, Area Code\_A415, Account Length, No of VMail Mesgs, Total Night Calls*** are used in the predictive classification tree model with only the selected features.

A picture containing calendar

Description automatically generated

the evaluation metrics for this decision tree classification model:

The Classification Tree Model **with** selected features accuracy score **is**: 0.7240

precision recall f1-score support

FALSE 0.91 0.75 0.82 850

TRUE 0.29 0.59 0.39 150

accuracy 0.72 1000

macro avg 0.60 0.67 0.61 1000

weighted avg 0.82 0.72 0.76 1000

The confusion matrix plot for the decision tree model with selected features:

Chart, treemap chart

Description automatically generated

Accuracy of this decision tree with selected features is 72.4%;

precision and recall are really low at 29% and 59%.

**Compare the two Decision Tree prediction models – baseline vs selected features:**

|  |  |  |
| --- | --- | --- |
|  | **Baseline Decision Tree Model** | **Selected Features Decision Tree Model** |
|  | ***Using all features*** | ***Using only selected features*** |
| **Accuracy** | 92.70% | 72.40% |
| **Precision** | 0.76 | 0.29 |
| **Recall** | 0.75 | 0.59 |

the decision classifier with all attributes more accurate with better performance metrics.

### Classification using Naïve Bayes (Unsupervised Learning)

**The Naïve Bayes baseline model with all features evaluation metrics:**

The Baseline Naive Bayes Model accuracy score **is**: 0.5790

precision recall f1-score support

FALSE 0.89 0.57 0.70 850

TRUE 0.20 0.61 0.30 150

accuracy 0.58 1000

macro avg 0.55 0.59 0.50 1000

weighted avg 0.79 0.58 0.64 1000

The confusion matrix plot:

Chart, treemap chart

Description automatically generated

**The Naïve Bayes baseline model with selected features evaluation metrics:**

The Naive Bayes Model **with** selected features accuracy score **is**: 0.3870

precision recall f1-score support

FALSE 0.91 0.31 0.46 850

TRUE 0.18 0.83 0.29 150

accuracy 0.39 1000

macro avg 0.54 0.57 0.38 1000

weighted avg 0.80 0.39 0.44 1000

The confusion matrix plot :

Chart, treemap chart

Description automatically generated

|  |  |  |
| --- | --- | --- |
|  | **Baseline Naïve Bayes Model** | **Selected Features Naïve Bayes Model** |
|  | ***Using all features*** | ***Using only selected features*** |
| **Accuracy** | 57.90% | 38.70% |
| **Precision** | 0.20 | 0.18 |
| **Recall** | 0.61 | 0.83 |

the Naïve Bayes predictive model with all attributes is more accurate with better performance metrics than the model with only selected features

Compare the two classification techniques - Decision Tree vs Naïve Bayes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Baseline Decision Tree Model** | **Selected Features Decision Tree Model** | **Baseline Naïve Bayes Model** | **Selected Features Naïve Bayes Model** |
|  | ***Using all features*** | ***Using only selected features*** | ***Using all features*** | ***Using only selected features*** |
| **Accuracy** | 92.70% | 72.40% | 57.90% | 38.70% |
| **Precision** | 0.76 | 0.29 | 0.20 | 0.18 |
| **Recall** | 0.75 | 0.59 | 0.61 | 0.83 |

# Conclusions and Recommendations

Best Model:

decision tree classification model with an accuracy score of 92.7%. a precision rate of 0.76.

Include the others Eric and Raymond added from SAS: Already mentioned in ppt.