**Author: Priyank Soni**

# Data Exploration and Understanding

1. **Load and Inspect Data:**

We have two files: train.csv (Features), train\_churn\_labels.csv (Labels).

There is total 230 features (columns) and 5000 samples. We observed that there are too many columns with many missing values, some of them are having more than 45000 values missing so it may not contribute much to your analysis and could be dropped, we decided threshold of 95% to drop these columns.

Now we have dataset size: 50000x77 (numerical: 43, categorical: 34)

1. **Exploratory Data Analysis (EDA):**

Summary statistics for numerical features:

| **Statistic** | **Var6** | **Var7** | **Var13** | **Var21** | **Var22** |
| --- | --- | --- | --- | --- | --- |
| count | 44471.00 | 44461.00 | 44461.00 | 44471.00 | 44991.00 |
| mean | 1326.44 | 6.81 | 1249.69 | 234.52 | 290.25 |
| std | 2685.69 | 6.33 | 2794.96 | 565.57 | 704.49 |
| min | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 25% | 518.00 | 0.00 | 0.00 | 112.00 | 135.00 |
| 50% | 861.00 | 7.00 | 232.00 | 144.00 | 180.00 |
| 75% | 1428.00 | 7.00 | 1604.00 | 228.00 | 285.00 |
| max | 131761.00 | 140.00 | 197872.00 | 36272.00 | 45340.00 |

**Key Insights:**

1. **Presence of Outliers:** The large difference between the mean and maximum values for Var6, Var13, Var21, and Var22 suggests the presence of outliers or extreme values. The positive skewness, as indicated by the mean being higher than the median, supports this.
2. **Skewed Distributions:** Most variables seem to have a positively skewed distribution, especially Var6 and Var13, as indicated by the mean being significantly higher than the median (**50th percentile** or **second quartile (Q2)**).
3. **Variability:** Var6 and Var13 have high standard deviations, indicating more variability in the data. These variables might have a wide range of values, as evidenced by the difference between their minimum and maximum values.
4. **Concentration of Values:** Var7 shows less variability, with a low standard deviation and relatively small differences between the 25th, 50th, and 75th percentiles.
5. **Data Imbalance:** Var22 has the highest count, indicating fewer missing values compared to other variables, while Var7 and Var13 have the lowest counts, suggesting some missing data.

Summary statistics for categorical features:

| **Statistic** | **Var192** | **Var193** | **Var194** | **Var195** | **Var196** | **Var197** | **Var198** | **Var199** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| count | 49,631 | 50,000 | 12,784 | 50,000 | 50,000 | 49,857 | 50,000 | 49,996 |
| unique | 361 | 51 | 3 | 23 | 4 | 225 | 4,291 | 5,073 |
| top | qFpmfo8zhV | RO12 | SEuy | taul | 1K8T | 0Xwj | fhk21Ss | r83\_sZi |
| freq | 385 | 35,964 | 12,567 | 47,958 | 49,550 | 4,629 | 4,441 | 955 |

**Key Insights:**

1. **Missing Data:**
   * **Var194** shows a substantial amount of missing data, which might require imputation or exclusion depending on the analysis goals.
   * Other variables are mostly complete, which is beneficial for analysis.
2. **Skewed Distributions:**
   * **Var193, Var195,** and **Var196** are highly skewed towards one category, which might imply a need for transformation or encoding if used in modeling.
   * This skewness suggests these variables could be important features in distinguishing between different classes if the dominant category holds specific meaning.
3. **High Cardinality:**
   * **Var198** and **Var199** have high cardinality, indicating a large number of distinct categories. This can add complexity to models, requiring strategies like one-hot encoding, frequency encoding, or embedding techniques.
4. **Low Variability:**
   * **Var196** might not contribute much to models due to its low variability, as nearly all entries are the same. It could potentially be a candidate for removal or require feature engineering to extract any useful information.
5. **Binary or Ternary Variables:**
   * **Var194** is essentially a binary or ternary feature, which could simplify certain modeling tasks but also might imply more about the nature of the missing data.

**Data Visualization:**

Histogram for numerical feature:



**Key Observations:**

1. **Distribution Shape**:
   * **Right-Skewed Distribution**: The histogram is heavily skewed to the right. This means that the majority of the data points are concentrated on the left side of the histogram, and there are a few extreme values or outliers on the right side. This kind of distribution is common in datasets where the values cannot be negative, and a few observations are significantly larger than the rest.
2. **Concentration of Data**:
   * Most data points are clustered towards the lower end of the scale, with a high frequency of observations near the smallest values (likely close to zero). This is evident from the high peak at the beginning of the histogram.
3. **Presence of Outliers**:
   * There are some significant outliers or higher values extending far into the right side of the histogram. These are the long tails that stretch across the x-axis with low frequencies, indicating a few extreme values that are much larger than the rest.

**Potential Causes and Implications:**

* **Causes of Right Skewness**:
  + **Natural Phenomena**: Many natural phenomena, such as income distribution, housing prices, and population sizes, tend to be right-skewed due to the presence of outliers.
  + **Data Collection**: Skewness might result from the way data was collected or limitations in the measurement process.
* **Implications**:
  + **Statistical Analysis**: Right-skewed data can affect statistical analyses, particularly those that assume normality (e.g., linear regression, t-tests). Transformation techniques, such as logarithmic or square root transformations, might be necessary to normalize the data.
  + **Decision Making**: Understanding the skewness is essential for making informed decisions. It might be necessary to focus on the median rather than the mean, as the mean can be affected by extreme values.

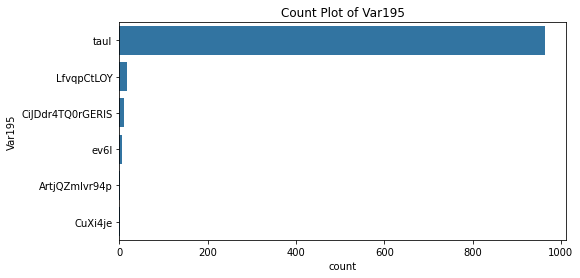
Box plot for numerical feature:



**Key Observations:**

* The box plot indicates a right-skewed distribution with many outliers.
* The data has a small IQR but a large range due to outliers.
* The presence of these outliers and the skewed distribution suggests that the data might benefit from transformations (like log transformation) or further cleansing for analysis.
* If this variable is used in a model, the effect of these outliers should be carefully considered, possibly using methods like robust scaling or transforming the data to reduce skewness.

This is the count plot for a categorical feature:



**Key Observations:**

* This count plot highlights a heavily dominant category in the dataset.
* It is essential to consider the implications of this distribution, especially if the feature is critical in predictive modeling or decision-making processes.
* Further steps might include data balancing or deeper exploration to understand the significance of this categorical feature.

# B. Data Preprocessing

For numerical features we are imputing missing values with the median

StandardScaler from sklearn.preprocessing module is used to standardize the features of a dataset. It transforms each feature by removing the mean and scaling it to have unit variance. This is often referred to as **z-score normalization** or **standardization.**

**For categorical variables we are imputing missing values with the most frequent value and we are using one hot encoding because** categorical features seems to have no ordinal relationship.

We are dropping highly correlated features with correlation thresholds = 0.8

# C. Model Development

Since we are dealing with high dimensional data with non-linear relationships between features and labels.

We start with Random Forests because it uses bagging and feature randomness, which leads to a diverse set of classifiers and reduces model variance. The ensemble nature of Random Forests averages out the noise, leading to a more generalizable model. It is less prone to overfitting because they aggregate multiple trees, which reduces the likelihood of learning the noise in the data.

We implemented a pipeline that includes preprocessing, feature selection, model training, and evaluation.

For handling class imbalance, we are using StratifiedKFold cross validation. It ensures each fold has the same proportion of class labels.

# D. Model Evaluation

To evaluate a classification model's performance on an imbalanced dataset, accuracy alone is not sufficient. Instead, metrics like F1 score, ROC AUC, precision, and recall provide more insight into how well the model handles class imbalance.

Cross-Validation Scores:

Accuracy Scores: [0.93333333 0.93333333 0.93333333 0.93333333 0.93333333 0.93333333

0.93333333 0.93333333 0.93333333 0.93333333]

Mean Accuracy: 0.9333333333333333

Standard Deviation Accuracy: 0.0

F1 Scores: [0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]

Mean F1: 0.0

Standard Deviation F1: 0.0

Roc\_auc Scores: [0.5 0.28571429 0.78571429 0.28571429 0.17857143 0.21428571

0.07142857 0.10714286 0.07142857 0.28571429]

Mean Roc\_auc: 0.2785714285714286

Standard Deviation Roc\_auc: 0.20873770280289222

Precision Scores: [0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]

Mean Precision: 0.0

Standard Deviation Precision: 0.0

Recall Scores: [0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]

Mean Recall: 0.0

Standard Deviation Recall: 0.0

# E. Future Improvements

**Handling class imbalance**

We can handle class imbalance because Imbalanced classes can lead to models that are biased toward the majority class, thereby affecting the overall performance of the model. These are the methods:

**1.Resampling Techniques**

* **Oversampling the Minority Class:** Increase the number of instances in the minority class by duplicating existing examples or generating new ones.
* **Undersampling the Majority Class:** Reduce the number of instances in the majority class to balance the dataset.

**2. Synthetic Data Generation**

* **SMOTE (Synthetic Minority Over-sampling Technique):** Create synthetic samples for the minority class.

**3. Using Class Weights**

* **Adjusting Class Weights:** Assign higher weights to the minority class so that the classifier pays more attention to it.

**4. Ensemble Methods**

* **Using Balanced Random Forests or EasyEnsemble:** These methods incorporate sampling strategies within the ensemble framework.

**Dry different encodings for categorical variables:**

We can try target encoding which replaces categorical values with the mean of the target variable for each category, which is useful for high-cardinality features.

We can try with models other than randomforest:

* 1. **Logistic Regression with Class Weights**
  2. **Support Vector Machines (SVM) with Class Weights**
  3. **Gradient Boosting Trees (XGBoost, LightGBM, CatBoost)**