

CUSTOMER CHURN

AADRIKA SINGH

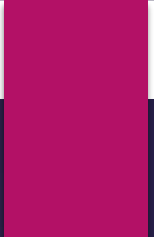
WHAT IS CUSTOMER CHURN?

- ▶ customers stopping the use of a service
- ▶ switching to a competitor service
- ▶ switching to a lower-tier experience in the service
- ▶ reducing engagement with the service

WHY WOULD A CUSTOMER CHURN?

In the case of telecommunication industries :

- ▶ number of service providers available
- ▶ perceived frequent service disruptions
- ▶ poor customer service experiences
- ▶ better offers from other competing carriers



“ It is always more difficult and expensive to acquire a new customer than it is to retain a current paying customer ”

WHY IS CUSTOMER CHURN IMPORTANT FOR TELECOM INDUSTRIES?

HOW TO REDUCE CUSTOMER CHURN?

- ▶ predict in advance which customers are going to churn through churn analysis
 - focus on a specific group rather than using retention strategies on every customer
 - big customer base and cannot afford to spend much time and money for it
- ▶ know which marketing actions will have the greatest retention impact on each particular customer

CUSTOMER CHURN ANALYSIS

The data analysis was carried out in the following sequence:

- ▶ Dataset Acquisition
- ▶ Data Wrangling
- ▶ Exploratory Data Analysis
- ▶ Machine Learning
- ▶ Model Evaluation
- ▶ Model selection

1. DATASET ACQUISITION

The data for this project comes from the data made available as a part of KDD Cup 2009: Customer Relationship prediction

LIMITATIONS:

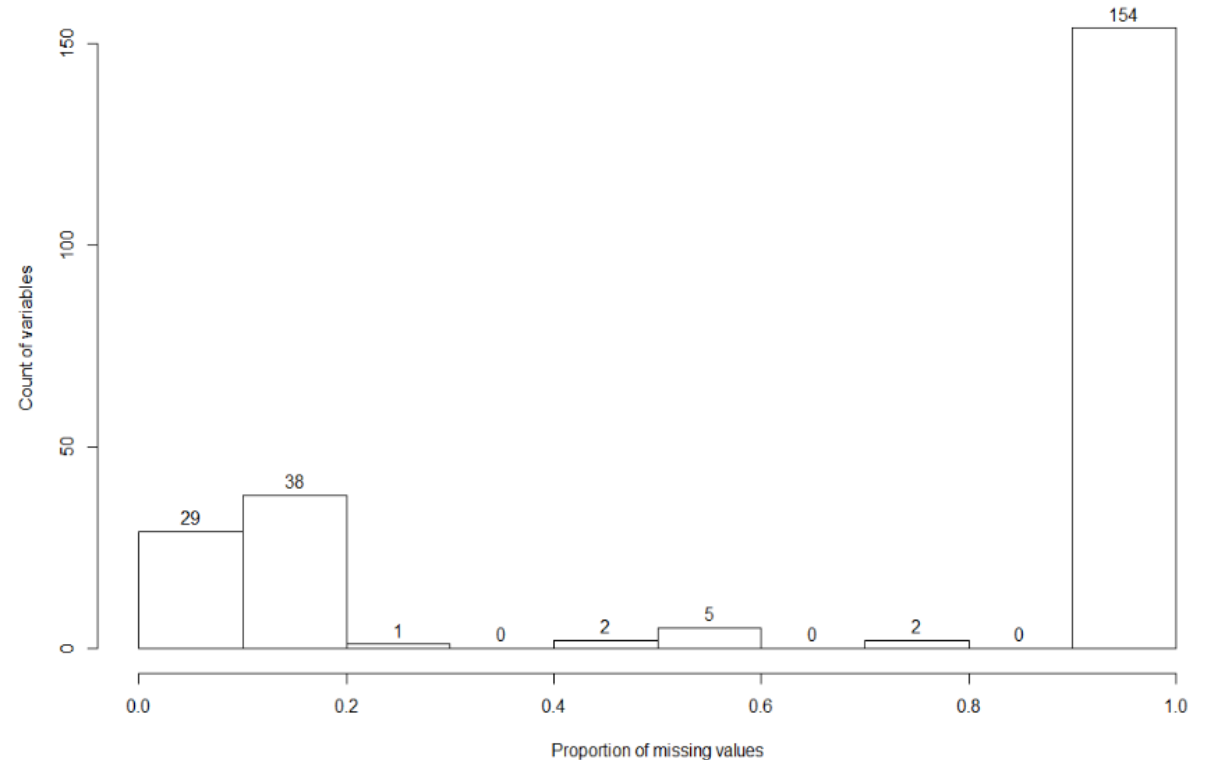
- ▶ Variables named as Var1, Var2, etc., which makes it difficult to identify which customer attribute is being looked at.
- ▶ Similarly, the information contained in categorical variables seem to be coded, again making it to difficult to relate to the kind of information contained in these fields.
- ▶ The dimensionality of the dataset is high (230 variables with 50000 observations), making it difficult to run certain transformations on the dataset, with the available infrastructural capacity.

2. DATA WRANGLING

The dataset contains a lot of missing values:

- columns with more than 20% of the data missing were removed
- variables with near-zero variance were removed
- rest of the missing values were imputed (process for *substituting* missing data)

Histogram of the proportion of missing values

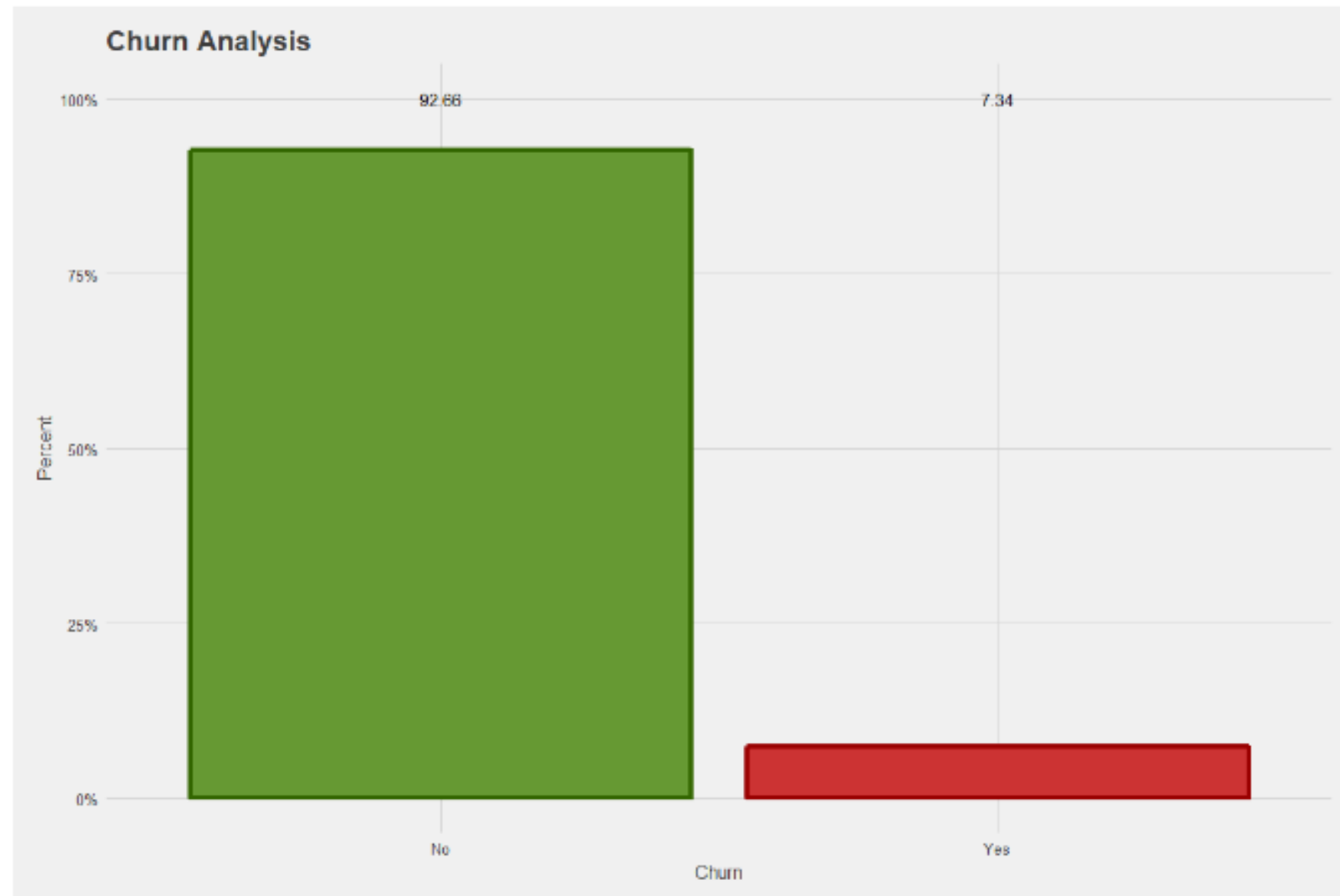


3. EXPLORATORY DATA ANALYSIS

Exploratory data analysis was performed, and important characteristics and statistical properties of the dataset were summarized and visualized.

Some examples of this analysis are contained in the following slides:

Distribution of people who churned



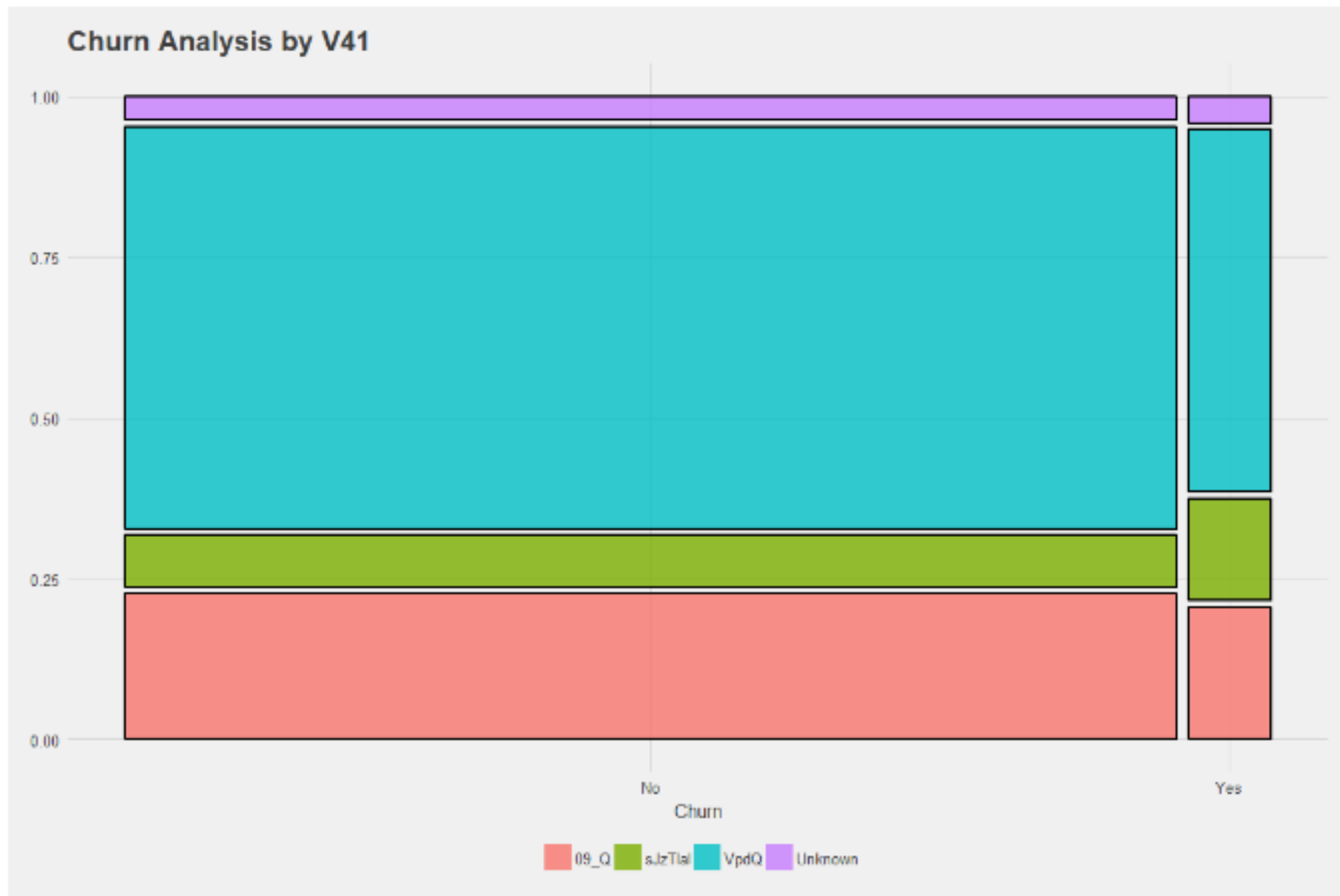
Finding: The percentage of people churning is much lower than the percentage of people not churning, which in turn, also implies, that the dataset is highly *imbalanced*.

Relationship between V49 and whether the customer churned or not



Finding: Based on this mosaic plot, it can be said that, given a customer churns, the V49 value is more likely to be **UYBR** than **cJvF**. Also, if V49 value is **Unknown**, then the customer is more likely to churn.

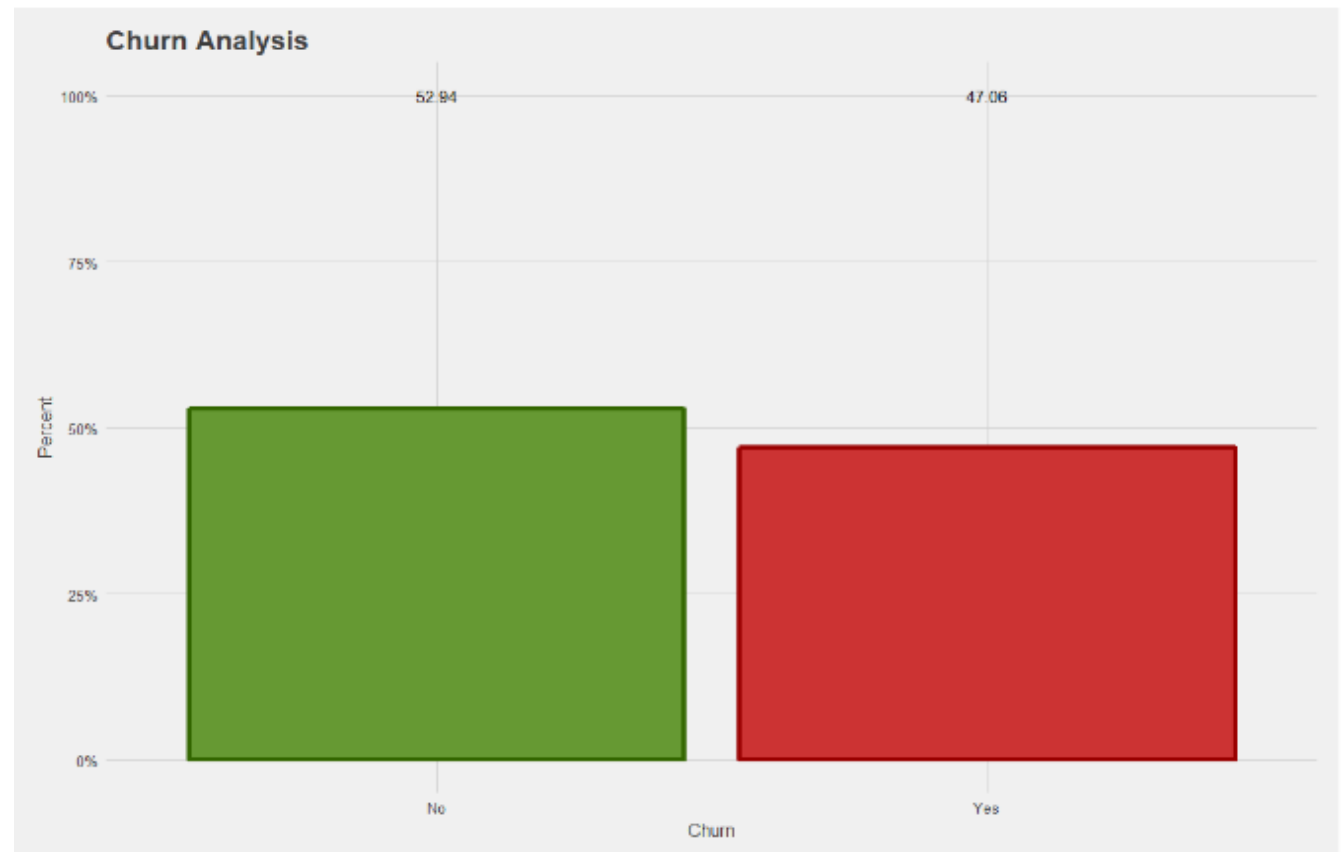
Relationship between V41 and whether the customer churned or not



Finding: If V41 value is **sjzTlal**, then the customer is more likely to churn.

4. MACHINE LEARNING

Since the dataset was highly imbalanced, oversampling technique was used for balancing the two classes, for proper data analysis.



4. MACHINE LEARNING

Four predictive models were fit to the dataset, and were compared against each other, on the grounds of several metrics.



5. MODEL EVALUATION & SELECTION

The model which had the highest precision, i.e. the highest confidence in its predictions, was selected.

According to the data, 1102 customers churned, whereas the model nearly predicted 50% (567 out of 1102) correctly.

ACTUAL	PREDICTION	
	Will Not Churn	Will Churn
	Didn't Churn	Churned
Didn't Churn	8753	5145
Churned	535	567

CONCLUSION

The results obtained have been optimized within the scope of this project; however, the model is not very confident about the predictions it makes about the customers, concluding that the analysis had a negative outcome

RECOMMENDATIONS

- ▶ Discussed in the exploratory data analysis section, there are several attributes and factors that might cause a customer to churn. For instance, if V41 value is sJzTlal, then the customer is more likely to churn. This can cause the business to identify the problem with this scenario of V41, and take effective actions to reduce the churn among customers.
- ▶ Even though the model has low precision value, it has a moderate recall value, suggesting that even though the confidence with which the model predicts the high risk customers is low, it still reduces the customer retention efforts by directing them solely towards a smaller customer base, rather than focusing on the entire customer base.