upGrad

#LifeKoKaroLift

Post-Graduate Diploma in ML/AI

26-09-2020

upGrad

Course : Machine Learning

Lecture On : Telecom Churn Case Study

Instructor : Manish Kumar

Agenda

- Problem Statement Overview
- What is churn?
- EDA
- Solution Sub-components

Customer Churn

upGrad

On average it's cheaper and easier to retain existing customers than to acquire new ones. An increasing number of top executives are refocusing their efforts on toward customer retention. Let's break down Telecom customer churn.

In the telecom industry, customers are able to choose from multiple service providers and actively switch from one operator to another. In this highly competitive market, the telecommunications industry experiences an average of 15-25% annual churn rate.

To reduce customer churn, telecom companies need to predict which customers are at high risk of churn.

In this project, you will analyze customer-level data of a leading telecom firm, build predictive models to identify customers at high risk of churn and identify the main indicators of churn.

There are two main models of payment in the telecom industry -

- **Postpaid model**, when customers want to switch to another operator, they usually inform the existing operator to terminate the services, and you directly know that this is an instance of churn.

- **Prepaid Model**, customers who want to switch to another network can simply stop using the services without any notice, and it is hard to know whether someone has actually churned or is simply not using the services temporarily

Customer Churn

upGrad

There are various ways to define churn,

- **Revenue-based churn**: Customers who have not utilized any revenue-generating facilities such as mobile internet, outgoing calls, SMS etc. over a given period of time.

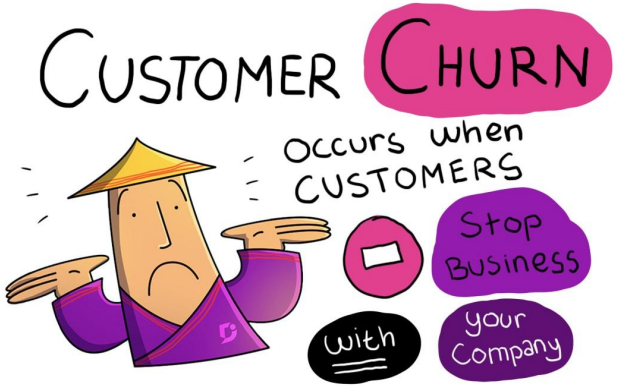
- **Usage-based churn**: Customers who have not done any usage, either incoming or outgoing - in terms of calls, internet etc. over a period of time.

In this project, you will use the **usage-based definition** to define churn.

Derive Churn

9th Month is our Churn Phase. Usage-based churn

1. Calculate total incoming and outgoing minutes of usage
2. Calculate 2g and 3g data consumption
3. Create churn variable: those who have not used either calls or internet in the month of September are customers who have churned
4. Check Churn percentage.
5. Delete columns that belong to the churn month

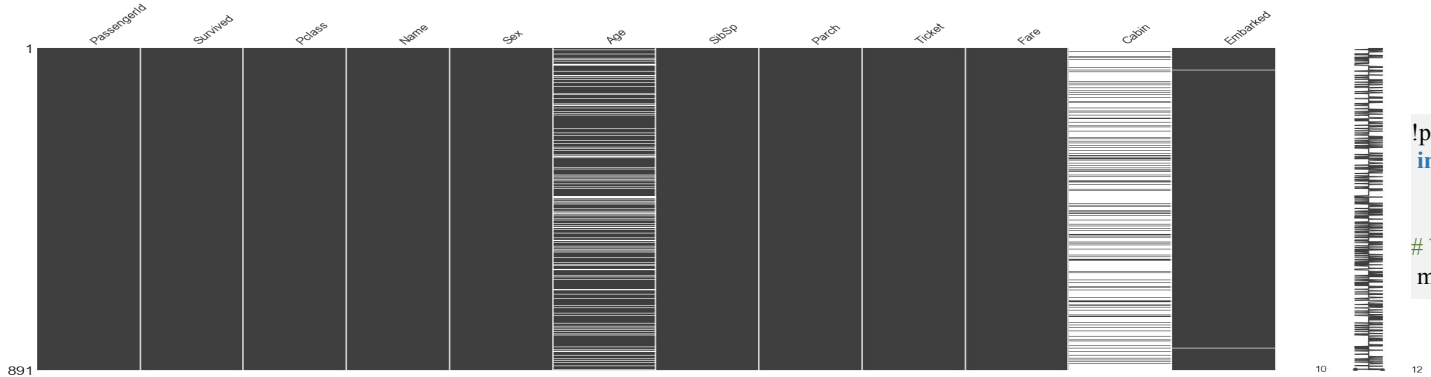


Missing Value Treatment

upGrad

1. **Data Extraction**: It is possible that there are problems with extraction process. In such cases, we should double-check for correct data with data guardians. Some hashing procedures can also be used to make sure data extraction is correct. Errors at data extraction stage are typically easy to find and can be corrected easily as well.

2. **Data collection**: These errors occur at time of data collection and are harder to correct. They can be categorized in four types:



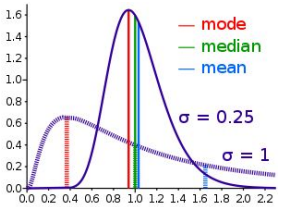

```
!pip install missingno
import missingno as msno


# Visualize missing values as a matrix
msno.matrix(df)
```

Methods to treat missing values

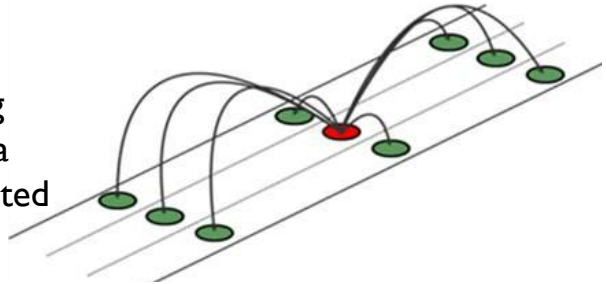
upGrad

Data – deletion Deletion methods are used when the nature of missing data is "**Missing completely at random**" or we have good amount of data and the data loss would be really low ,else non-random missing values can bias the model output



Mean/ Mode/ Median Imputation Mean / Mode / Median imputation is one of the most frequently used methods. It consists of replacing the missing data for a given attribute by the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable.

Prediction Model we create a predictive model to estimate values that will substitute the missing data. In this case, we divide our data set into two sets: One set with no missing values for the variable and another one with missing values. First data set become training data set of the model while second data set with missing values is test data set and variable with missing values is treated as target variable.



https://www.kaggle.com/athi94/investigating-imputation-methods

High Value Customers

upGrad

Filter high-value customers(HVC)

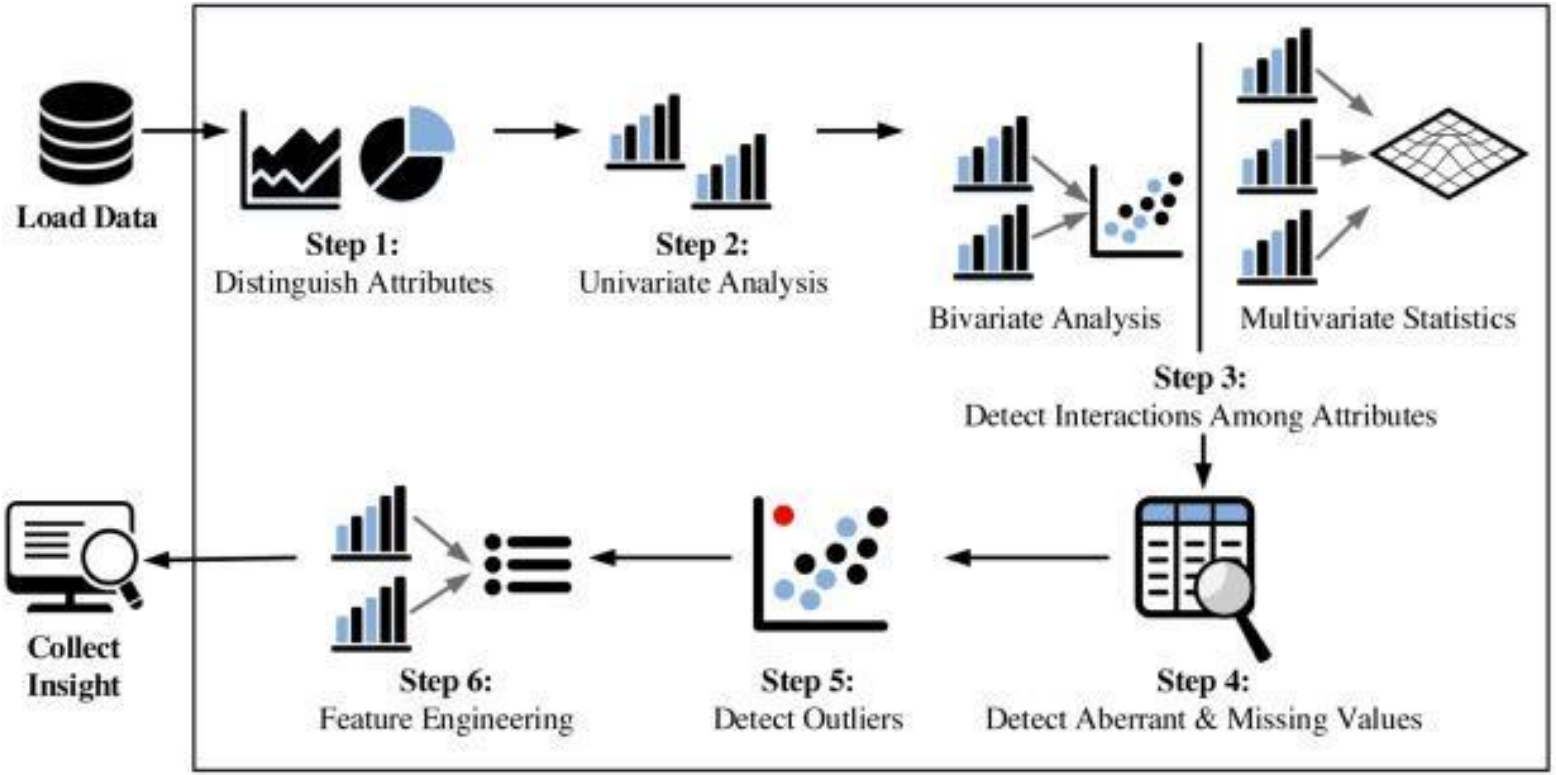
1. Calculate total data recharge amount
2. calculate total recharge amount
   a. call recharge amount(total_rech_amt) + data recharge amount

3. Calculate average recharge done by customer in June and July
4. Look at the 70th percentile recharge amount

5. Retain only those customers who have recharged their mobiles with more than or equal to 70th percentile amount

Fundamental steps of EDA process

upGrad

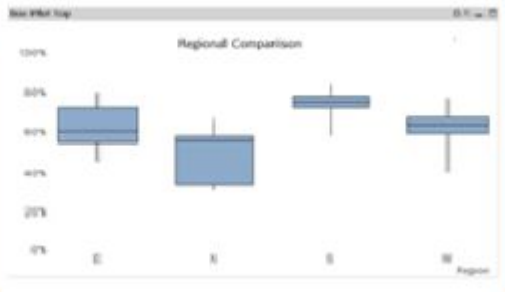


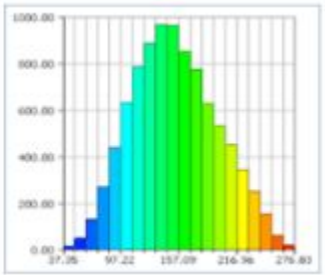
Univariate Analysis

upGrad

At this stage, we explore variables one by one. Method to perform univariate analysis will depend on whether the variable type is categorical or continuous.

Central Tendency	Measure of Dispersion	Visualization Methods
Mean	Range	Histogram
Median	Quartile	Box Plot
Mode	IQR	
Min	Variance	
Max	Standard Deviation	
	Skewness and Kurtosis	

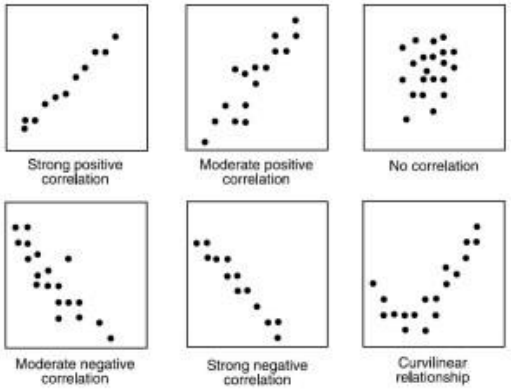




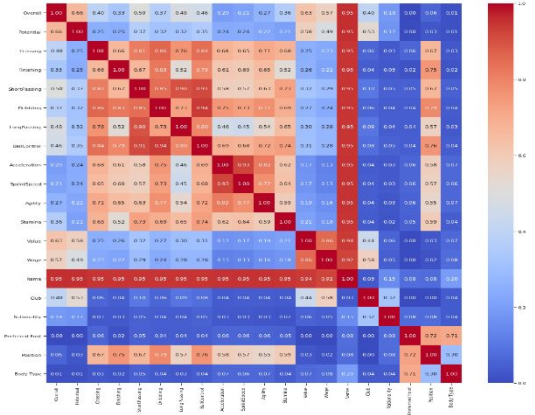
Continuous vs continuous

upGrad

While doing bi-variate analysis between two continuous variables, we should look at scatter plots. It is a nifty way to find out the relationship between two variables. The pattern of scatter plot indicates the relationship between variables. The relationship can be linear or non-linear.



Scatter Plots



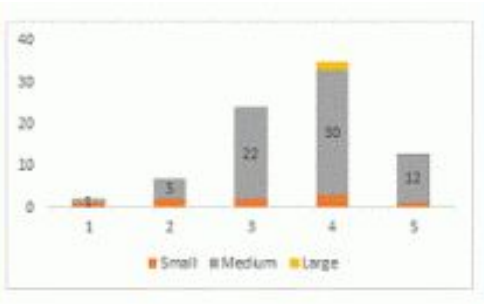
Correlation Heat Maps

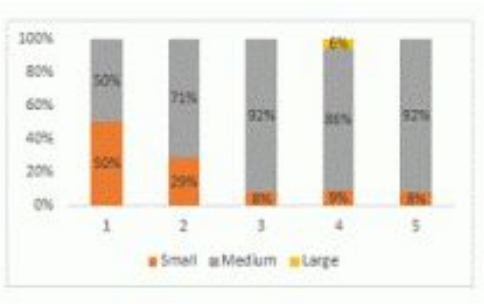
Categorical & Categorical

upGrad

Frequency / Row Pct	Product Category 1	2	3	4	5	Total
Small	1 11.11	2 22.22	2 22.22	3 33.33	1 11.11	9
Medium	1 1.43	5 7.14	22 31.43	30 42.86	12 17.14	70
Large	0 0.00	0 0.00	0 0.00	2 100.00	0 0.00	2
Total	2	7	24	35	13	81

Frequency Missing = 77

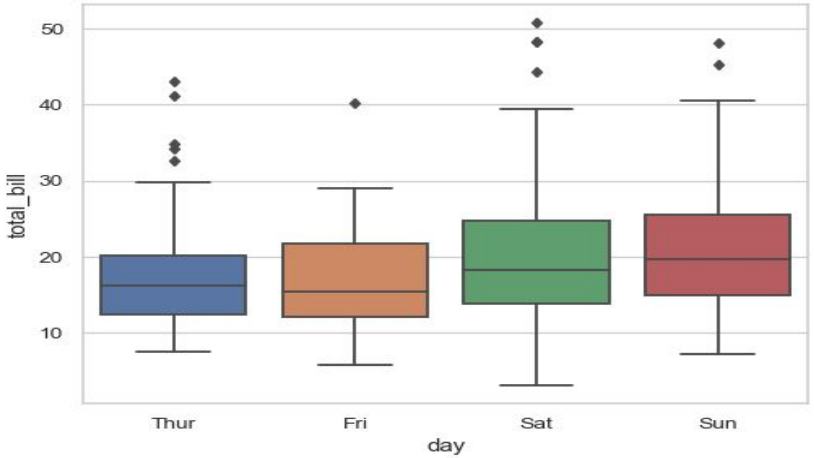




Two-way table: We can start analyzing the relationship by creating a two-way table of count and count%. The rows represents the category of one variable and the columns represent the categories of the other variable. We show count or count% of observations available in each combination of row and column categories.

Stacked Column Chart: This method is more of a visual form of Two-way table.

Categorical & continuous



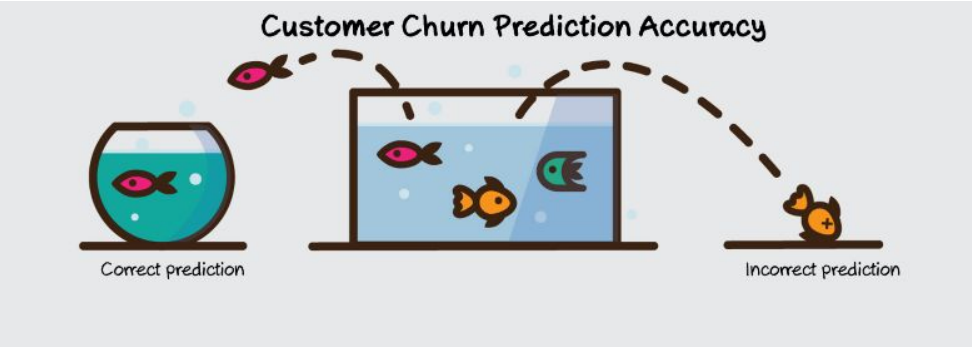
While exploring relation between categorical and continuous variables, we can draw box plots for each level of categorical variables

upGrad

Modelling

1. Need two models, one with good predictions and other with good interpretability.
2. Use PCA to reduce the variables.
3. You need to handle the imbalance class.

https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18



upGrad

#LifeKoKaroLift

Thank You!