Problem Statement

Company Information:

A telecom company called ‘Firm X’ is a leading telecommunications provider in the country. The company earns most of its revenue by providing internet services. Based on the past and current customer information, the company has maintained a database containing personal/demographic information, the services availed by a customer and the expense information related to each customer.

Problem Statement:

You are working for the telecom company ‘Firm X’. It has a customer base set across the country. In a city ‘Y’, which is a significant revenue base for the company, due to heavy marketing and promotion schemes by other companies, your company is losing customers i.e. the **customers are churning**. Whether a customer will churn or not will depend on data from the following three buckets:

1. Demographic Information
2. Services Availed by the customer
3. Overall Expenses

The data are provided in **three separate data files** given at the end of the page**.** The aim is to automate the process of predicting if a customer would churn or not and to find the factors affecting the churn. The collated **data dictionary** for the variables in the 3 data frames is given below:

S.No.	Variable Name	Meaning
1.	CustomerID	The unique ID of each customer
2.	Gender	The gender of a person
3.	SeniorCitizen	Whether a customer can be classified as a senior citizen.
4.	Partner	If a customer is married/ in a live-in relationship.
5.	Dependents	If a customer has dependents (children/ retired parents)
6.	Tenure	The time for which a customer has been using the service.
7.	PhoneService	Whether a customer has a landline phone service along with the internet service.
8.	MultipleLines	Whether a customer has multiple lines of internet connectivity.
9.	InternetService	The type of internet services chosen by the customer.
10.	OnlineSecurity	Specifies if a customer has online security.
11.	OnlineBackup	Specifies if a customer has online backup.
12.	DeviceProtection	Specifies if a customer has opted for device protection.
13.	TechSupport	Whether a customer has opted for tech support of not.
14.	StreamingTV	Whether a customer has an option of TV streaming.
15.	StreamingMovies	Whether a customer has an option of Movie streaming.
16.	Contract	The type of contract a customer has chosen.
17.	PaperlessBilling	Whether a customer has opted for paperless billing.
18.	PaymentMethod	Specifies the method by which bills are paid.
19.	MonthlyCharges	Specifies the money paid by a customer each month.
20.	TotalCharges	The total money paid by the customer to the company.
21.	Churn	This is the target variable which specifies if a customer has churned or not.

How to start the case study:

To solve any analytics problem, the Crisp-DM framework is to be followed. The chart below will guide you through the process of solving the case study:

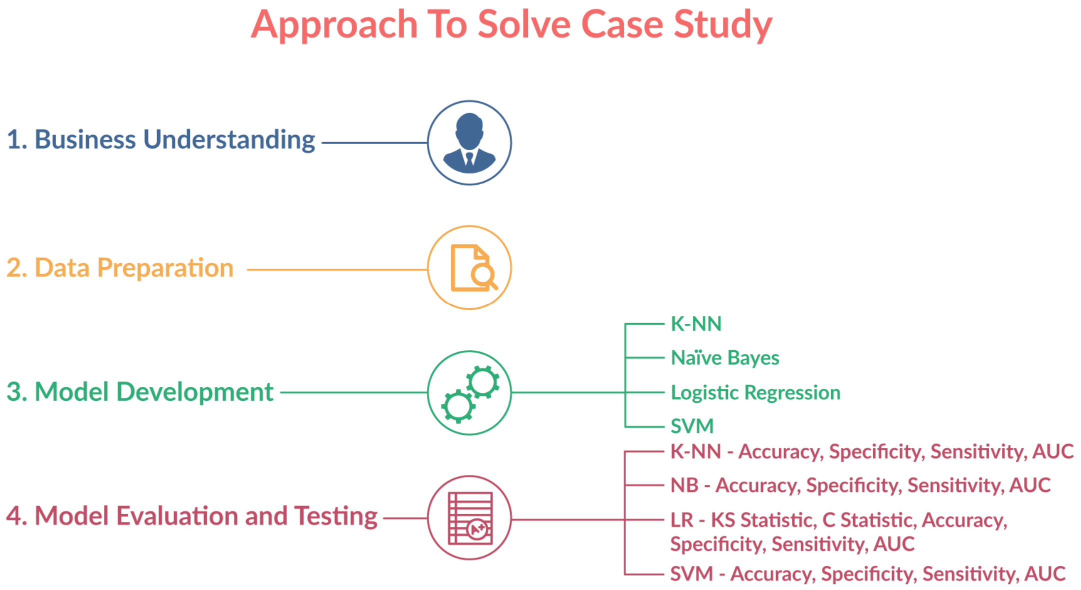
Approach To Solve Case Study
1. Business Understanding
2. Data Preparation
3. Model Development
K-NN
Naïve Bayes
Logistic Regression
SVM
4. Model Evaluation and Testing
K-NN - Accuracy, Specificity, Sensitivity, AUC
NB - Accuracy, Specificity, Sensitivity, AUC
LR - KS Statistic, C Statistic, Accuracy, Specificity, Sensitivity, AUC
SVM - Accuracy, Specificity, Sensitivity, AUC

Figure 1: Case Study Flow

The goal of this case study:

You are required to develop predictive models using each of the 4 models namely K-NN, Naive Bayes, Logistic Regression and SVM.

Note: Wherever required, set the VIF threshold to 2.

Note:

Please make sure the below points are to be followed strictly for evaluation purpose:

- Store the collated dataset into “churn” object.

- Divide the dataset into 70:30 ratio and set seed to 100 for the reference. It should be renamed as “train” and “test” respectively.

Packages Required:

You have to install the below packages below starting this assignment.

- install.packages(“MASS”)
- install.packages(“car”)
- install.packages(“e1071”)
- install.packages(“ROCR”)
- install.packages(“caret”)

Checkpoints

1. Business Understanding

Customer churn can depend on a lot of internal and external factors. External factors are the ones you might have no control or information about, for example, the launch of Reliance Jio will lead to churn across all telecom companies and predicting the churn for such cases gets extremely difficult. However, Internal factors such as the demographic information, the number of connections taken by a customer, personal information, billing information, information on services availed etc. can be used to predict the churn of customers. In this case study, you only need to consider the internal factors.

2. Data Understanding

The data is provided in 3 different files in the previous segment. The file customer.csv contains the personal information of the customers. The file churn_data.csv contains the information related to the churn of customers along with their billing information. The third file contains the information related to the internet usage and all the services customers are using.

3. What is the Company’s Business Objective?

The company wants to understand the driving factors behind churn and wants to build a model which would predict future churn. The company can utilise this knowledge for churn prevention. Specifically, the company wants to determine which driver variables are having the most influence on the tendency of churning.

Your goal is divided into 5 main parts:

1. Data Understanding
2. Data Preparation
3. Model Building
4. Model Evaluation
5. Presentation of results

Problem Statement:

How do we approach the case study?

The entire case study is divided into 4 checkpoints to narrow down the problem statement. The checkpoints are interlinked with each other. Use the commented R file given below to write the code.

Checkpoint-1: Data Understanding and Preparation of Master File.

- Load the 3 files given in separate data frames.
- Collate the data together in one single file.

Checkpoint -2: Exploratory Data Analysis

- Make bar charts displaying the relationship between the target variable and various other features and report them.

Checkpoint -3: Data Preparation

- Perform de-duplication of data.
- Bring the data in the correct format
- Find the variables having missing values and impute them.
- Perform outlier treatment if necessary

Checkpoint 4: Model Building

1. **Model - K-NN:**
    - Data Prep:

        - Prepare data for K-NN (Don't forget to convert categorical variables to numeric)
    - Modelling and Evaluation.
        - Build a K-NN model using the optimal K
        - Report the performance metrics - Accuracy, Sensitivity, Specificity and AUC.
2. **Model - Naive Bayes:**
    - Data Prep:
        - Make sure that the variables have the correct data type.
    - Modelling:
        - Build the Naive Bayes model (Hint: use- NaiveBayes() command from e1071 package).
        - Report the performance metrics - Accuracy, Sensitivity, Specificity and AUC.
3. **Model -Logistic Regression:**
    - Data Prep:
        - Prepare the data for logistic regression. (Hint: Logistic regression takes numeric variables as inputs)
    - Modelling: Part 1: Keep the probability threshold as 0.3, 0.5 and 0.7 respectively.
        - Perform Variable Selection
        - Report the final logistic regression model.
        - Report the performance metrics - Accuracy, Sensitivity, Specificity and AUC, KS-Statistic and C-Statistic.
4. **Model - SVM:**
    - Data Prep:
        - Prepare the data for SVM. (Hint: SVM takes numeric variables as inputs)
    - Modelling

- Make the SVM model by finding the optimal value of ‘cost’ and report its performance metrics - Accuracy, Sensitivity, Specificity and AUC.

As a final result, report the best model for each algorithm along with the most impactful variables and its performance metrics (wherever required).

This would help us understand which variables, in fact, are affecting the tendency to churn.

Document template (Mandatory submission): Download the sample .docx file below. Also, download the commented R file which you have to use to write your codes and submit for evaluation.