Executive summary

Table of contents

1. Introduction

We deal with client profiles related data, from a famous telecom compagny. We perform an exploratory data analysis to help out our friend hired in the marketing department of the firm. We are interested in the lifetime value of loyal customers.

2. Exploratory Data Analysis

2.1 Data cleaning

In the framework of the data type analysis of each feature, we figure out that while we do not have NaN values, we do have empty cells in the TotalCharges feature.

As a result, we perform a replacement of cells consisting of one space, by NaN values, which we then filter out from data. We thus remove 11 rows with missing data.

As we focus on lifetime value of loyal customers, we do not consider those who leave the firm.

2.2 Feature Selection

Before starting the feature selection per se, we do a little of feature engineering:

- **Create a new variable 'MonthlyExpense' = 'TotalCharges'/'tenure'**: As we are intrested in the lifetime value of the contracts, we can compute the average monthly expense of each customer, then the lifetime value of the contracts could be extrapolate from the average lifetime of the contracts. Therefore the feature selection will consist in identifying which features have the most explanatory power regarding this new feature 'MonthlyExpense'. As we already suspect, by construction, some features such as 'TotalCharges', 'tenure' and 'MonthlyCharges' will certainly be highly correlated with 'MonthlyExpense' and could be discarded in the subsequent feature importance analysis in order to obtain a better resolution of the other features contribution to our dependent variable.
- **Transform the categorical variable into numerical values**: Indeed as we have seen during de preliminary data inspection, we have a lot of categorical feature that we will need to map to numerical value. When it is a binary value, e.g. yes/no or male/female we can directly replace it by 0 or I. When there is more possibility like for the 'contract' feature we will use the pd.factorize() function to establish a numerical mapping. Furthermore, notice that some feature present redundant information such as the 'No phone service' information that is already present in the 'PhoneService' feature and the 'No internet service' present in all the internet related services features that could be placed in a separated feature: 'InternetService' under the form of true or false (an additional column will be created for 'Fiber'->True/False). Proceeding this way most of the categorical feature will become binary except 'contract' and 'payement method'.

2.2.1 Correlation Of Feature

As mentioned before, by construction, we already expected a high correlation with 'TotalCharges' and 'MonthlyCharges'. Without taking them into account, the highest correlated features with 'MonthlyExpense' are:

- 'Fiber'
- 'Internet'
- 'StreamingMovies'
- 'StreamingTV'

2.2.2 Univariate Selection

We perform a serie of univariate regression of the dependent variable ('MonthlyExpense') on each individual feature to observe the individual effect of each of them.

Once again discarding 'TotalCharges' and 'MonthlyCharges', the top five F-score obtained in these univariate regression are for the following feature:

- 'Fiber'
- 'Internet'
- 'StreamingMovies'
- 'StreamingTV'
- 'MultipleLines'

Regarding their p-values all of them are statistically relevant.

2.2.3 Feature Importance

Feature importance gives a score for each feature of your data based on an inbuilt class that comes with Tree Based Classifiers, the higher the score more important or relevant is the feature towards the output variable. once again the top five is given by:

- 'Fiber'
- 'Internet'
- 'StreamingMovies'
- 'StreamingTV' == 'PhoneServices'

3. Conclusion

'Fiber', 'Internet', 'StreamingMovies' and 'StreamingTV' probably play a determinant role in the explanation of the 'MonthlyExpense' and a fortiori of the lifetime value of the contracts.

On the aggregate, data is clean. However, data types had to be specified.

As the main cashflow drivers of a telecom compagny are centered around their data plans, we understand that what drives a client's MonthlyExpsenses are the four services mentioned above. Additional services like OnlineBackup and OnlineSecurity, do not seem to drive the MonthlyExpenses.

From our heatmap, we had figured out that clients' profiles and therefore consumption, does not seem to be correlated with age or sex. One might be interested in varying the advertising strategy to see whether we can introduce a bias in the data, ceteris paribus.

In [1]:
```
import os

os.system('jupyter nbconvert --to html Executive_Summary.ipynb')
```

Out[1]:
```
0
```

In [ ]:
```

```