Data Mining 2018/2019 Report
Telco Customer Churn dataset

Andrea Pedrotti, mat: 569692
Lorenzo Taverniti, mat: 431197
Gurban Aliyev, mat: 589206

May 2019

Abstract

Data mining project for Data Mining 2018-2019 class at Pisa University.

1 Data Understanding

Telco Customer Churn dataset contains information concerning customers' contract gathered by *Telco* telecommunication company. Goal of the analysis is to gain insight into the reasons that lead customers to leave or stay with the company. The dataset is composed of 7043 entries and 21 attributes.

1.1 Data Semantics

The two tables below describe the attributes found in the dataset. They are split between numerical and categorical data.

Categorical Type	**Attribute Name**	**Domain**
Binary	Gender SeniorCitizen Partner Dependents PhoneService Churn	{True, False}
Nominal	CustomerID	Alphanumerical
	MultipleLines	{Yes, No, No Phone Service}
	InternetService	{DSL, Fiber, No}
	OnlineSecurity OnlineBackup DeviceProtection TechSupport StreamingTv StreamingMovies	{No, Yes, No Internet Service}
	PaymentMethod	{Electronic check, Mailed check, Bank transfer, Credit card}
Ordinal	Contract	{Month-to-month, 1year, 2year}

Table 1: Categorical Values table.

Eighteen of the attributes contained in the dataset are categorical values. Six of them are binary attributes, eleven are pure nominal values while the last one is ordinal.

1. **Gender** refers to the customer gender.
2. **SeniorCitizen** refers to whether the customer is a senior citizen or not.
3. **Partner** refers to whether the customer has a partner or not.
4. **Dependents** refers to whether the customer has dependents or not.

5. **PhoneService** refers to whether the customer has chosen to opt in for Telco's phone service or not.

6. **Churn** represents whether the customer has chosen stay with or leave the company

Data Type	Attribute Name	Domain
Discrete	Tenure	$[0, 72] \cap \mathbb{N}$
Continuous	MonthlyCharges	$[18.25, 118.75] \cap \mathbb{R}$
	TotalCharges	$[0.0, 8684.80] \cap \mathbb{R}$

Table 2: Numerical Values table.

Three attributes are numerical. Two of them are continuous while the last one is discrete.

1. **Tenure** represents number of months the customer has stayed with the company.

2. **MonthlyCharges** is the the monthly amount charged to the customer.

3. **TotalCharges** is the total amount charged to the customer.

1.2 Data Distribution Visualization

To visualize numerical attributes we both plotted the normal distribution with respect to churning (Figure 1 - first row), and the density-based distribution with respect to the two target-class values: churn no or churn yes (Figure 1 - second row).
Both *Tenure* and *MonthlyCharges* follow a bi-modal distribution. With respect to *Tenure* plot, we can observe that there is a really high amount of new customers, while between 20 and 60 months there is almost a flat distribution. The number of customer churning strongly decreases over time.
Concerning *MonthlyCharges* plot, most of the customers are charged a low amount of money. The monthly amount spent by the customer is correlated with the churning decision: the lower, the smaller the chance the customer will cancel his subscription. However, also customer who are charged the highest are less likely to change provider. With regards to *TotalCharges*, there is a peak of customer with lowest total charge which indicates the high amount of new customers with a 'budget' contract. As expected, with the increase of the values, the ratio of customer churning decreases: they are strongly loyalized customer both because they are long-time customers or because they are willing to spend a high amount of money to receive the best service (e.g. they opted in for all *premium services*).

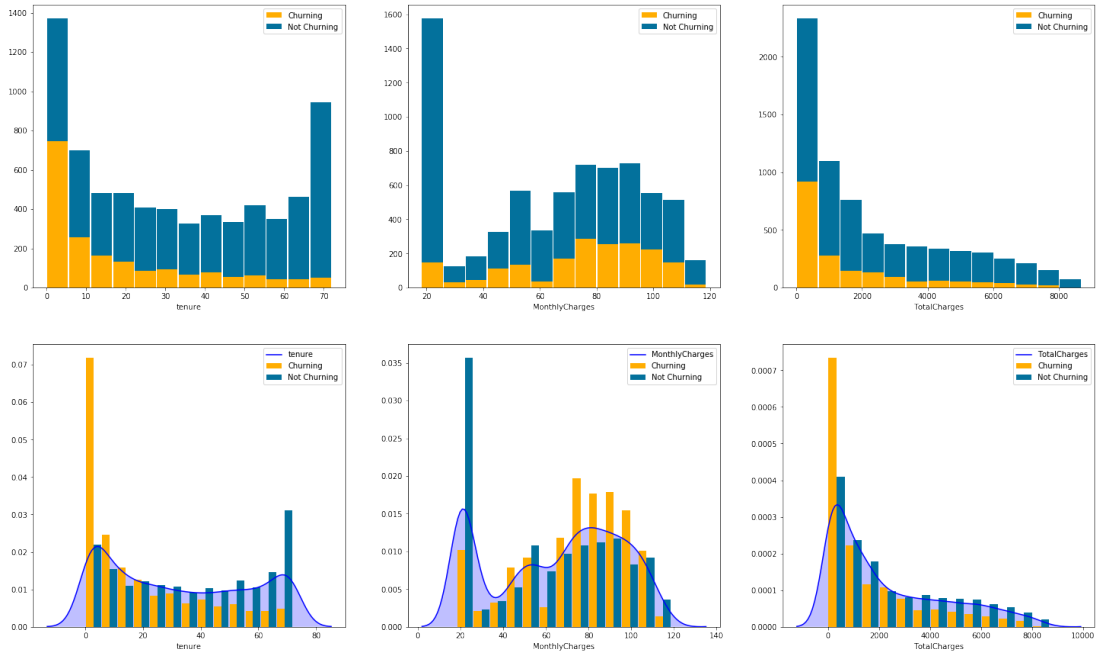

Figure 1: Frequency plot for continuous attributes.

Attributes	**Mean**	**Mode**	**Median**	**Std**
Tenure	32.37	1	29	24.56
MonthlyCharges	64.76	20.05	70.35	30.09
TotalCharges	2279.73	20.02	1394.55	2266.79

Table 3: Statistical measurement for numerical values .

We plotted every categorical attribute with respect to the churning decision. We report in figure 2 the results. Note that we have already tweaked certain attributes:

1. Original *InternetService* now contains only 0 and 1 values, marking customers who subscribed for internet services. Information about internet connection type now is stored into two new binary values: *DSL* and *Fiber*.

2. All internet premium services had been grouped into *PremiumServices* column.

3. Contract now has been split between short-term contract (month-to-month contracts) and long-term contract(one-year or two-year contracts.

4. *PaymentMethods* has been re-grouped by check payments and automatic ones. More on this in *Data Transformation* Section.

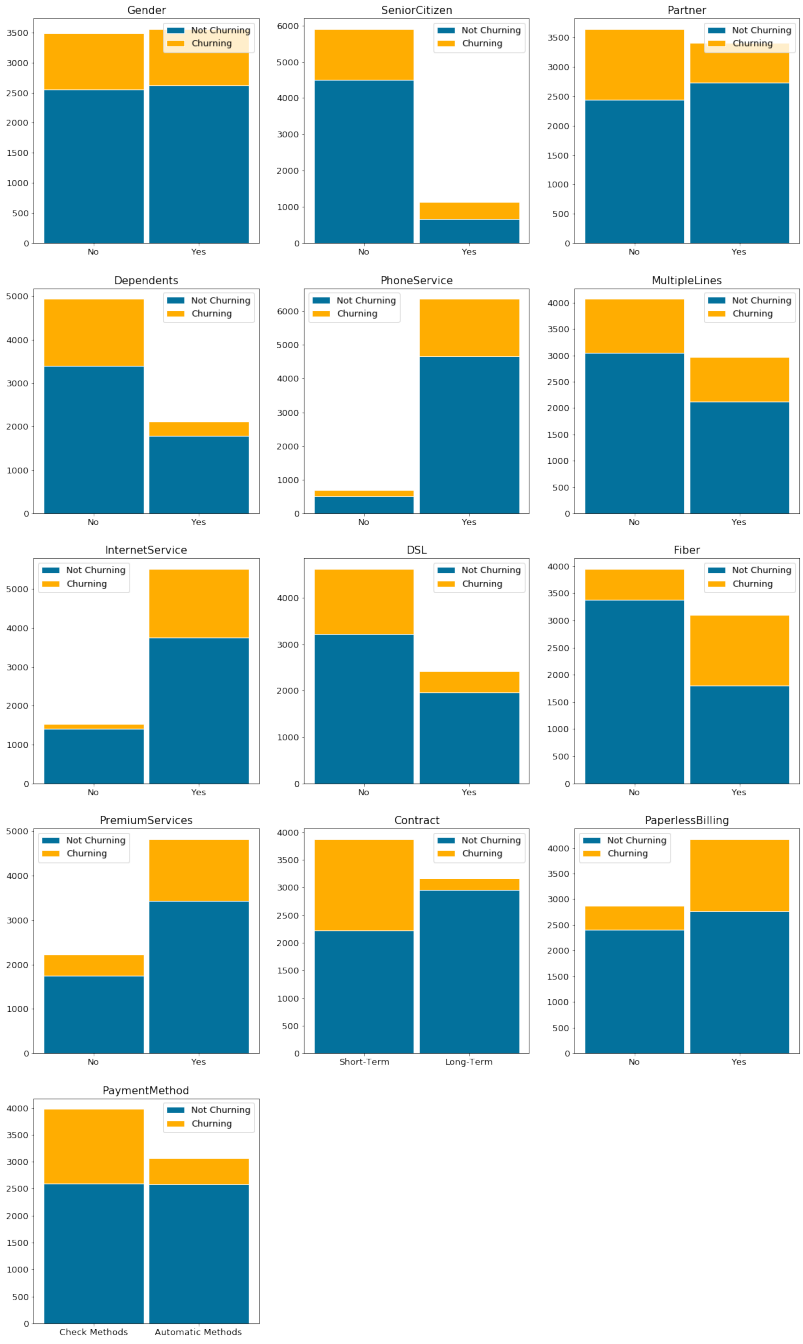

Figure 2: Frequency plot for categorical attributes.

From these plots we can make some early remarks:

1. There is an even distribution with respect to customers gender and partner attribute.
2. Almost all of the customers have opted in for PhoneService. This is less marked for *InternetServices*, however, still the big majority of customers is also subscribed for the service.
3. Customers with short-term contract are more likely to churn.

1.3 Data Quality

The quality of the given dataset has been assessed according to the following criteria.

1.3.1 Syntactic Accuracy

The dataset presents only just a minor syntactical error that have been immediately fixed: columns name are written in inconsistent form. They have been all renamed capitalizing the first letter.

1.3.2 Semantic Accuracy

Concerning semantic accuracy, The dataset has some small flaws. First of all, with respect to the attribute *TotalCharges* there are 11 entries with values ' ' (empty string). Furthermore, the original *InternetServices* condenses information about both the subscription to the service and its type. As already noted, we therefore split this information into two new columns.

1.3.3 Completeness and Missing Values

The dataset is complete. Just 11 values of the attributes *TotalCharges* are in fact missing. Therefore 99.85% of the data are already up to use. It is also easily noticeable that all these missing values are restricted to new customers who have subscribed to the company within the month of the dataset creation. Accordingly, we can already fill these values with zero.

1.3.4 Outliers

As depicted by the boxplots below, given the categorical skewness of the dataset, even before actually analyzing the data we could hypothesize the almost complete absence of outliers.

None of the three numerical values shows presence of outliers. It is important to note that the attribute *TotalCharges* is simply the product of *Tenure* and *MonthlyCharges*.

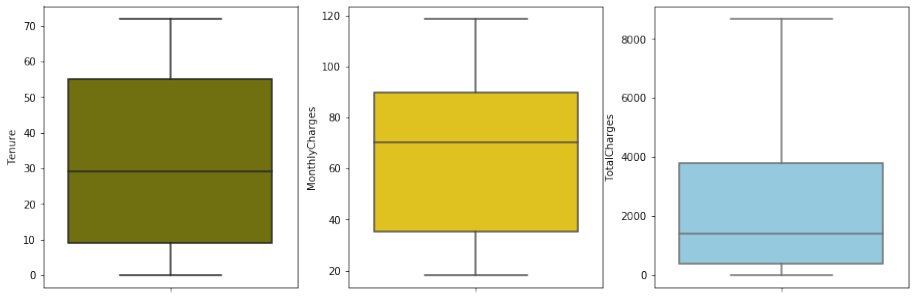

Figure 3: Boxplot for dataset numerical values.

1.4 Data Transformation

The dataset as given has no *Null-Object*. However, most of the values are represented in inconvenient forms.
In fact, almost all of the categorical values are stored as *string*. Therefore, we encoded all those values as *integer or float* values. The encoding for attributes *Gender, SeniorCitizen, Partner, Dependents, PhoneService, Churn* resulted in a binary categorization, while for attributes *MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, PaymentMethods* resulted in a nominal categorization, ranging from 0 to 1 for all except the latter, which ranges from 0 to 2. Lastly, the continuous attributes also had been normalized in range [0,1].

1.5 Data Correlation

We report the correlation matrix obtained on the processed dataset. As already noted thanks to the previous plots. We denote a strong correlation between type of contract and churning. There is also a logical high correlation between *Tenure* and *Contract*: as expected, this means that if a customer has a long-term contract he will probably stay with the company for a longer period. Concerning the target values *Churn*, both *Fiber* and *Tenure* reveal a negative correlation. This suggests that not-loyalized clients are more inclined to change company, if attracted by a competitor with a more convenient internet offer.
Lastly, we denote no correlation between personal attributes like *Gender* and *SeniorCitizen* and any other attribute.

2 Clustering Tasks

In this section we report the results obtained by applying clustering techniques: K-means, DBSCAN and Hierarchical Clustering. We have conducted the analysis on the three continuous vales present in the dataset. In fact, performing this clustering techniques on categorical attributes would have resulted in meaningless conclusion.

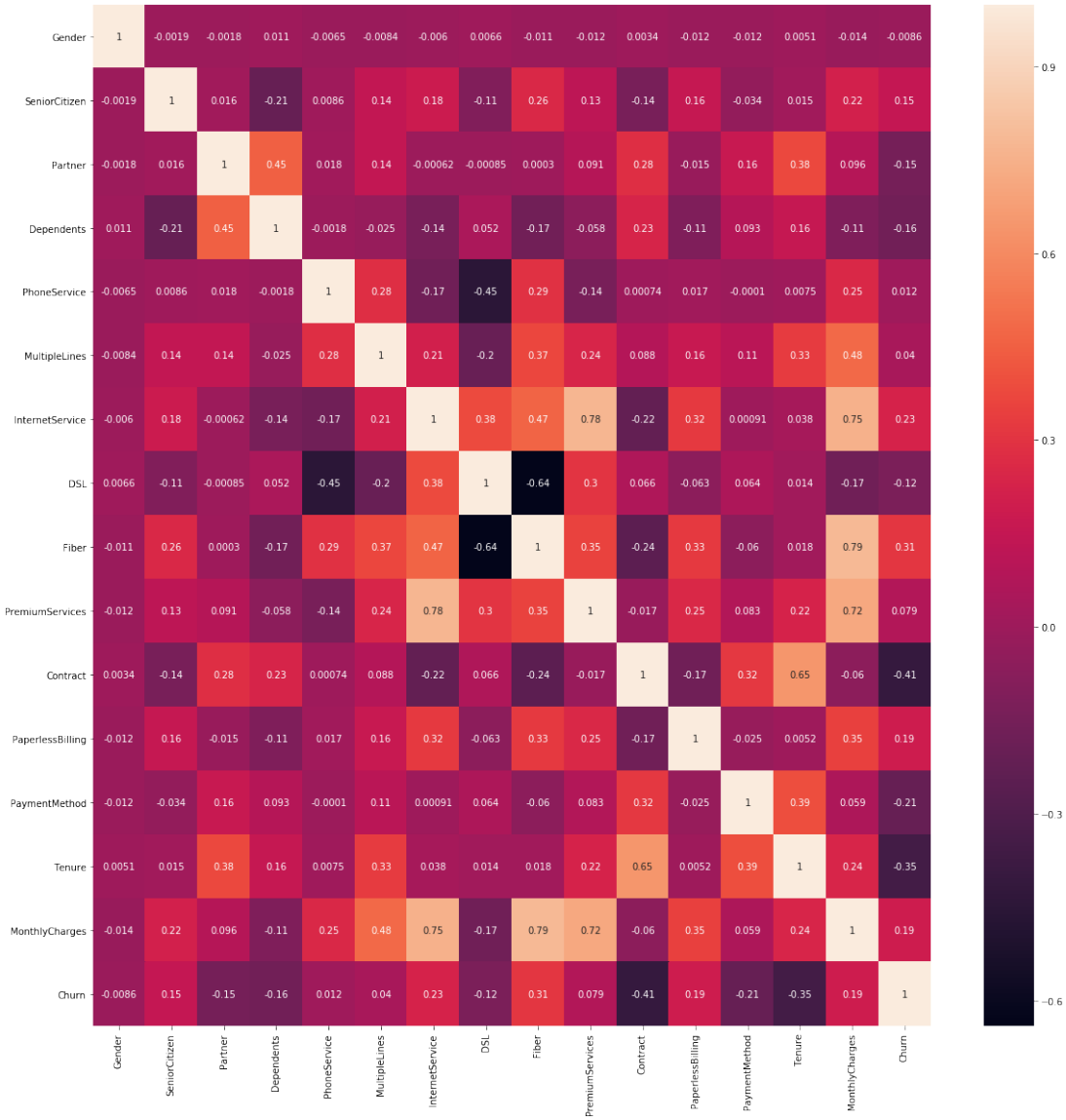

Figure 4: Correlation Matrix.

2.1 K-Means

To estimate the best number of cluster k, we have first performed K-means with k in range [1,100]. We then narrowed the range to range [2,15]. From the plot in figure 5 we can spot a knee in the SSE-curve at k equals to 5. Moreover, the silhouette plots reveals k equals to 9 as a good value, however we have chosen

the lower k because, by running the algorithm 10 times for a maximum of 100 iterations. At each iteration the initial centroids were chosen randomly. We report in table 4 some high-level statistics for the obtained clustering.

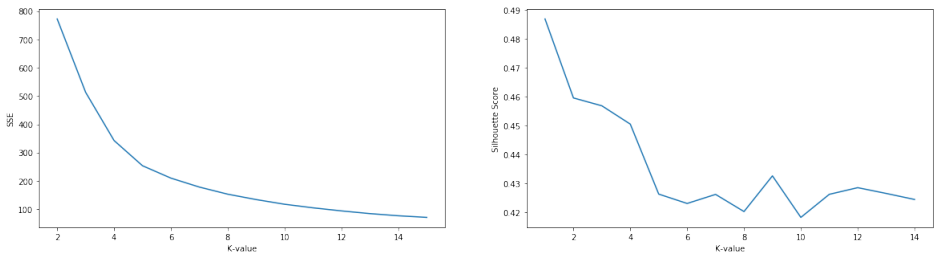

Figure 5: SSE and silhouette plots.

ID	Elements	ChurnRatio	Tenure Mode	MonthCh Mean
Cluster0	1701	54.08%	1 (267)	77.80
Cluster1	1501	12.59%	72 (286)	93.92
Cluster2	1003	4.28%	72 (76)	30.85
Cluster3	1580	23.86%	1 (364)	30.06
Cluster4	1258	27.02%	35 (58)	82.93

Table 4: High-level stats for k-means clustering, k=5.

To visualize the categorical composition of the clusters we plotted every attributes grouped by churning value. By looking at the histograms is easy to spot a prominent attributes characterizing the whole clustering. We report here the visualization just for cluster0 of k-means in order to avoid flooding the report with images even though they are a simply and immediate way to read the dataset.

Cluster0 is characterized by a small majority of churning customers (1: 920, 0: 781). Customers in this cluster are almost all subscribed for *internetServices, phoneServices* and have a short-term contract. Electronic check and Mailed check are the favorites payment methods. Tenure is generally low, ranging from 0 to 5 while *monthlyCharges* are mostly in a medium-high range [70,80].

Cluster1 is composed of almost only not-churning customers (0: 1312, 1: 189). Customers here are characterized by partner, premium services subscription, multiple line service and long-term contract. They are largely charged for low monthly amounts and are mostly long-time clients. This customer profile is generally the less likely client to churn. They are mostly long-time customers [60,70] with high-end contract [80,115].

Figure 6: Categorical distribution K-means Cluster0.

Cluster2 is also characterized by a strong majority of non-churning customers (0: 960, 1: 43). Clients in this group are not interested in internet service. Still, those who are, they all have DSL connection and tend to prefer automatic payment methods, long-term contracts. They generally avoid premium internet services. They are mostly long-time customer with low amount of monthly charges [10,30].

Cluster3 is another non-churning inclined cluster (0: 1203, 1: 377). Clients here are mostly single and prefer short-term contract. This cluster is similar to the previous one (cluster2) in terms of services. However, Customers from this cluster prefer short-term contract causing the cluster churning rate to be higher. They are mostly new customers [0,5] with low amount of monthly charges [10,30].

Cluster4, the last identified cluster, is again not-churning inclined (0: 918, 1: 340). Clients belonging to this cluster are interested in internet and inter-

net premium services and they tend to prefer Fiber connection and short-term contracts. They are mostly long-time customer [60,70] with mid-range monthly charges [50, 70].

In figure 7 we report the visualization of k-means clustering. On the z axis data are split between non-churning (0) and churning (1). By looking at the projection of each cluster on the z axis we can thus visually assess the quality of the clustering with respect to the target value *Churn*

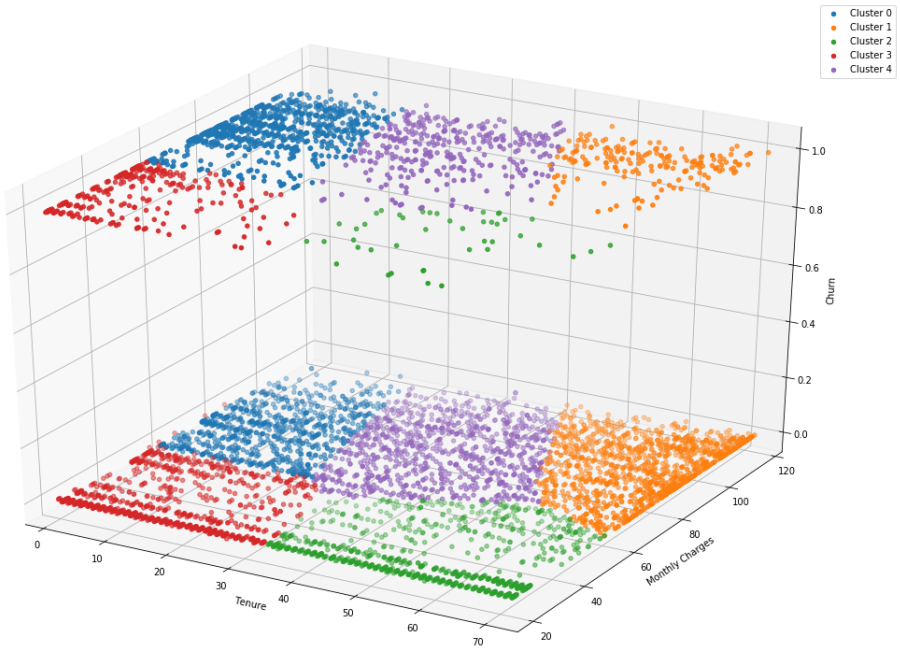

Figure 7: K-means clustering visualized in 3-dimension.

2.2 DBSCAN

In order to estimate the best parameters we have run the algorithm with epsilon ranging from 0 to 1. We then narrowed the research to the range [0.001,0.200] identifying as best epsilon 0.05728282828282828. Concerning the minimum number of samples in a neighborhood for a point to be considered as a core point, we identified 80 as optimal parameter. We also tried to obtain an equal number of cluster with respect to k-means in order to be able to compare more easily the cluster consistency between algorithms. To do this, we have also compared the clusters with *Gephi* by plotting over the k-means cluster the DBSCAN labels. We report the results in the following pages of this section.

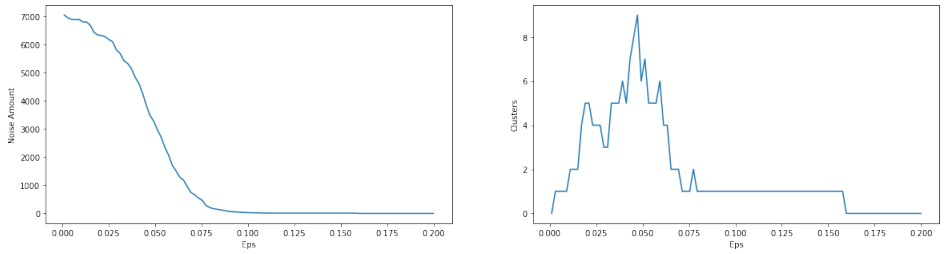

Figure 8: Noise Amount and number of cluster per epsilon.

ID	Elements	ChurnRatio	Tenure Mode	MonthCh Mean
Cluster0	1704	11.27%	1 (267)	21.88
Cluster1	2304	51.05%	1 (396)	74.68
Cluster2	788	10.79%	72 (247)	99.25
Cluster3	104	13.46%	24 (21)	54.38
Cluster4	76	23.69%	56 (16)	98.11

Table 5: High-level stats for DBSCAN clustering.

We have applied the same method illustrated for k-means to analyze DBSCAN clusters. We limit to report the a sample of continuous attributes visualization omitted for the sake of brevity in previous section.

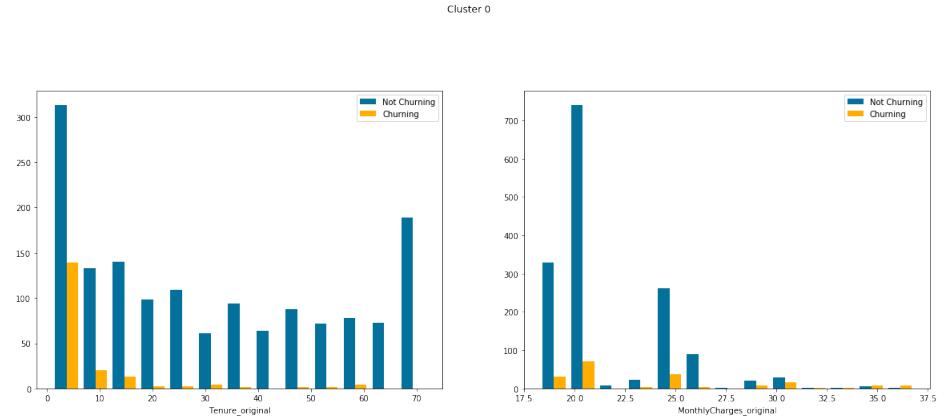

Figure 9: Histograms for continuous values DBSCAN Cluster0.

Cluster0 is non-churning cluster (0: 1512, 1: 192)of customers not subscribed to internet service. Those who are, they opted for DSL connection type and no internet premium services. Mostly made of new customers with low monthly charge amount.

Cluster1 is lightly skewed churning customers (0: 1128, 1: 1176) interested in internet service and Telco's premium services. Clients in this cluster signed short-term contract and prefer non-automatic payment methods with paperless billing. Cluster1 is composed mostly by new and medium time customers [0,10] with medium-high charge contracts [70,100].

Cluster2 is largely made out of non-churning customers (0: 703, 1: 85). Customers belonging to this clusters have internet subscription with fiber connection and long-term contracts. They prefer automatic payment methods. Cluster2 gathers long-term customers [70,72] with high monthly charges contracts [100-120]

Cluster3 is a small cluster largely made out of non-churning customers (0: 90, 1: 14). Customers in this cluster opted in for internet service and internet premium services even though they chose a DSL internet connection type. Short and long-term contracts are balances, as well as payment methods and paperless billing. The cluster is built around medium-time customers [22,28] with medium-charge contracts [50,60].

Cluster4 is another small cluster made out of non-churning customers (0: 58, 1: 18). This cluster resembles cluster3 with the main difference in the most common connection type. Customers are long-term clients [50,60] with high monthly charge contracts [95,105].

2.3 Hierarchical clustering

Concerning hierarchical clustering, we analyzed the same dataset used for k-means and DBSCAN. As distance metric we kept euclidean, as for the other clustering tasks. We evaluated both the *complete linkage* and *single-linkage*. As expected, complete-linkage showed the best result, as single-linkage is too prone to chaining, resulting in the creation of a single huge cluster (of 6984 entries). We report in figure 11 the dendrogram obtained for complete-clustering. For complete-linkage we chose to to cut it at 0.8, obtaining 5 different clusters in order to compare the clustering quality with respect to the other clustering techniques. By looking at the plot we note, however, that the range [0.52,0.75], could have been another good splitting point. Pertaining to single-linkage, we report the resulting dendrograms even though it carries no useful information. Therefore, we have limited the further comparison to complete-linkage algorithm.

2.4 Comparison

After labelling each datapoint with its own three cluster label, we used *Gephi* to create a network based on k-means clustering (central network in figure 12). We then colored each datapoint based on DBSCAN label (left network) and

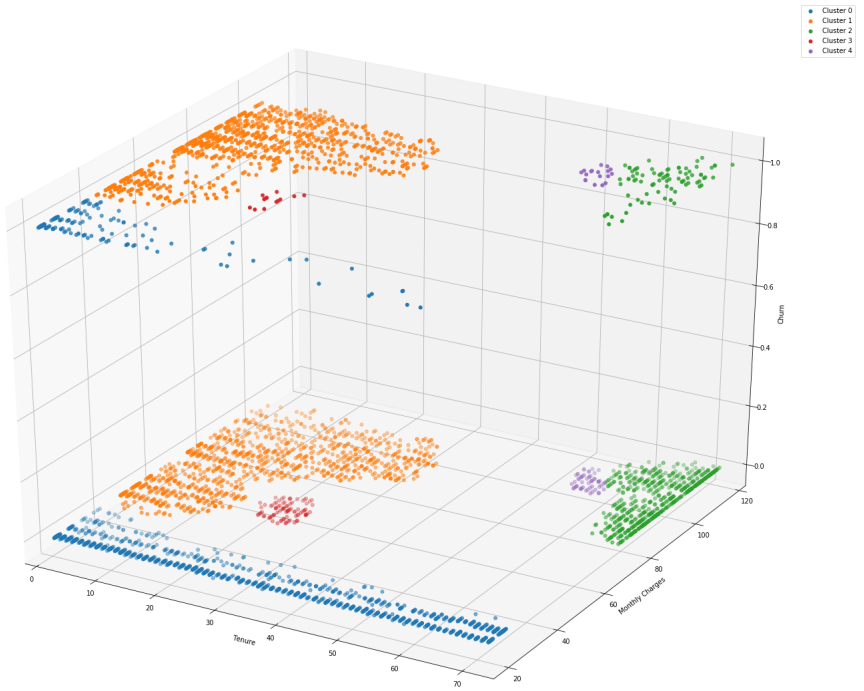

Figure 10: 3-dimension plot DBSCAN clusters.

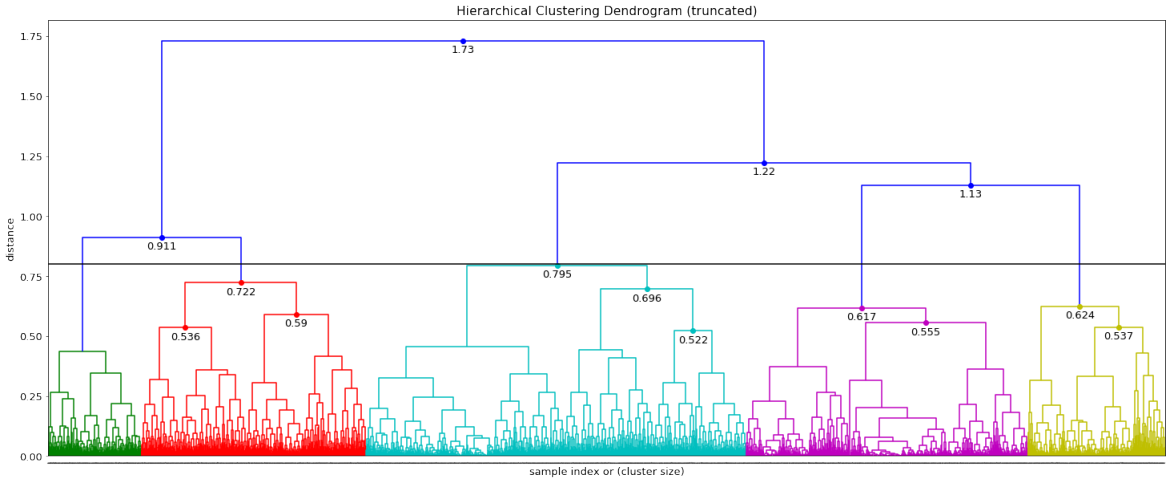

Figure 11: Complete-linkage, cut at 0.8.

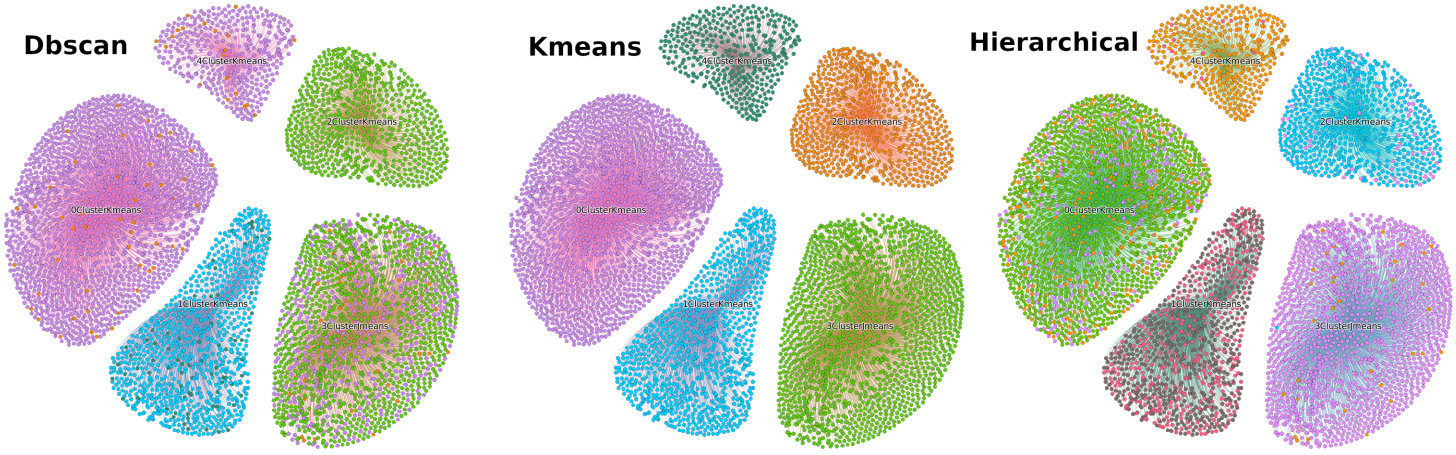

Figure 12: Comparison of clustering quality.

hierarchical (right network). We have also used this technique to assess each clustering quality with respect to relevant attributes. Again, for the sake of brevity we avoid reporting all of them and limit to showcase k-means clustering with respect to *Churn* attribute (figure 13).

We report in table 6 the different silhouette scores obtained.

Algorithm	**Silhouette**
K-means	0.45
DBSCAN	0.17
Hierarchical - complete	0.38

Table 6: Silhouette Scores.

3 Association Rules Mining Tasks

3.1 Frequent Items

We have extracted maximal pattern by setting minimum support to 20%.The identified itemsets are straightforward and in line with the results obtained both in data understanding and clustering tasks. We omit any commentary on these section and limit to report the 5 most frequent ones.

3.2 Association Rules

We have mined association rules by applying the *Apriori algorithm*. In order to extract the most meaningful rules, after testing it on the whole dataset, we have

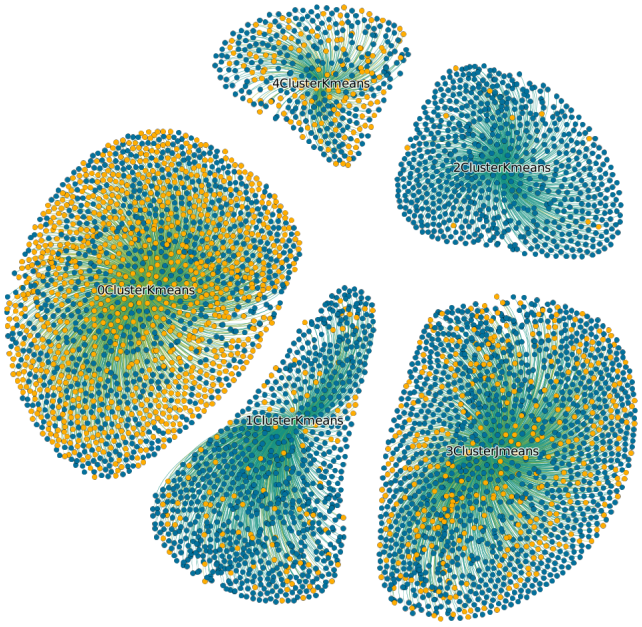

Figure 13: K-means cluster quality wrt Churn - Yellow: churning, Blue: not-churning.

Abs Support	**Itemset**
2088	('1_Fiber', '1_pLBil', '0_DSL', '1_premServ', '1_intSer', '1_phSer')
1921	'('1_pLBil', '1_premServ', '0_churn', 1_intSer', '1_phSer')
1832	('1_contr', '0_Fiber', '0_churn', '1_phSer')
1821	('1_contr', '0_DSL', '0_churn', '1_phSer')
1789	('1_DSL', '0_Fiber', '1_premServ', '0_churn', '1_intSer')

Table 7: Top 5 maximal pattern.

decided to reduce the number of attributes by keeping only those with a higher correlation with the target attribute *Churn*. By doing so, we have reduced the dataset to eleven attributes. The continuous attributes have been converted to categorical attributes by binning in 13 categories according to Sturges' rule.
After some initial test, we have noted that to extract rules concerning 1 as churn value, we had to lower the min_support parameter. This is due to the skewness

of the dataset with respect to the target value (0: 5174, 1: 1869). With this consideration in mind, we decided to extract general rules with a min_support parameter equal to 20, however, when extracting rules targeting churning, we lowered the parameter to 1. We first report the rule with highest lift score for each target attribute identified (min_sup = 20).

Target	**Lift**	**Sup**	**Rule**
No internetService	4.61	21.25%	('0_monCh', '0_premServ', '0_Fiber', '0_DSL', '1_phSer')
0/12 monthlyCharges	4.38	21.25%	'(0_intSer', '0_premServ', '0_Fiber', '0_DSL', '1_phSer')
No premiumServices	3.17	21.25%	('0_intSer', '0_monCh', '0_Fiber', '0_DSL', '1_phSer')
Yes DSL	2.91	25.40%	('0_Fiber', '1_premServ', '0_churn', '1_intSer')
Yes Fiber	2.27	21.41%	('0_contr', '0_paMe', '0_DSL', '1_intSer', '1_phSer')
No Fiber	1.78	21.25%	('0_intSer', '0_monCh', '0_premServ', '0_DSL', '1_phSer')
No long-term Contract	1.62	17.90%	('1_churn', '0_DSL', '1_phSer')
No DSL	1.52	21.25%	('0_intSer', '0_monCh', '0_premServ', '0_Fiber', '1_phSer')
Yes premiumService	1.44	23.58%	('1_contr', '0_churn', '1_intSer', '1_phSer')
No Churn	1.31	26.01%	('1_contr', '0_Fiber', '1_phSer')
Yes internetService	1.27	25.40%	('1_DSL', '0_Fiber', '1_premServ', '0_churn')
Yes phoneService	1.10	21.25%	('0_intSer', '0_monCh', '0_premServ', '0_Fiber', '0_DSL')

Table 8: Highest lift scoring rules.

Rules obtained are once again as expected. We limit to note how the rule targeting *0_churn* remarks the tendency to not-churn for customer with DSL connections. As we will remark in the conclusion section this goes to underline how the Fiber connection service could be the most profitable service to focus in order to lower the number of customer cancel their contract. We report in table 9 the most interesting rules obtained by lowering the minimum support to 1. By doing so we were able to extract also rules with respect to the churning decision.

Once again, rules in table 9 are far from reliable, but we did anyway to assess how much we had to lower the parameter in order to extract at least one rule concerning *1_Churn*.
Finally, we have built a rule-based classifier and assessed its accuracy by pre-

Target	Lift	Rule
Yes Churn	4.46	('7_monCh', '0_tenure', '1_Fiber', '0_contr', '1_pLBil', '0_DSL', '1_intSer', '1_phSer')
12/12 Tenure	6.19	('12_monCh', '1_paMe', '1_Fiber', '1_pLBil' '0_DSL', '1_premServ', '0_churn', '1_intSer', '1_phSer')

Table 9: Rules obtained with min_sup equals to 1.

dicting target value *Churn*. The classifier was fed the top 5 rules extracted from the a training set composed by 80% of the original dataset. Performances was then assessed on the remaining data. We report in table 10 the obtained results.

Metric	Training Set	Test Set
Accuracy	63.23%	63.55%
Precision	41.21%	41.25%
Recall	91.74%	89.55%
F1-score	56.87%	57.11%

Table 10: Evaluation for rule-based classifier.

4 Classification Tasks

Regarding classification, we applied *DecisionTree* classifier and *RandomForest* classifier. We also have used, for the Decision Tree model a grid search based on this parameter:

- Criterion: [entropy, gini]
- max depth:[1-30]
- min impurity decrease:[1*exp-6, 5*exp-6,1*exp-7,5*exp-6]

The dataset has been divided into *Training Set* and *Test Set*, respectively in 2/3 and 1/3 of the dataset, preserving the proportion of the attribute *Churn* in the original dataset. The best results obtained with a cross-validation on 10 folds, are summarized in the following tables. Table 11 shows the results of the model obtained with this parameters:

- Criterion: [gini], max depth:[5], min impurity decrease:[1*exp-6]

Instead, regarding the *RandomForest* classifier the best result, obtained through a grid search, is reported in Table 12 obtained with this parameters:

- Criterion: [gini], max depth:None, n estimators = 150, bootstrap = True, max features = 'sqrt'

Metric	Training Set	Test Set
Accuracy	79.62%	78.79%
Precision	61.39%	61.33%
Recall	61.68%	57.96%
F1-score	61.54%	59.65%

Table 11: Evaluation for Decision Tree classifier.

Metric	Training Set	Test Set
Accuracy	86.87%	79.68%
Precision	81.93%	66.22%
Recall	64.52%	51.74%
F1-score	72.17%	57.75%

Table 12: Evaluation for RandomForest classifier.

The problem that affects this dataset is the strong imbalance in favor of Not Churning people. In fact we can observe how in the results obtained in table 12 the lowest value is always that of Recall, that is, how many True instances we were able to correctly predict. This is caused by a high number of noise points and by a low population of Churn.

In figure 14 we have reported the decision tree obtained from the best model. We can observe how people who have a long-term contract are led to not leave; while those who have a short contract and internet with fiber tend to leave. Those with a DSL internet contract tend not to change; and those with high MontlyCharge and low tenure tend to change. While there is a good group of people that even having a high MonthlyCharge does not change.

The substantial difference between the two models developed is the different choice of important features. In fact we can observe from the figures 14-15 that the first model chooses as important features the type of contract, the type of internet line and the number of months of subscription. While the second one prefers the monthly cost of the subscription and the number of months of contract.

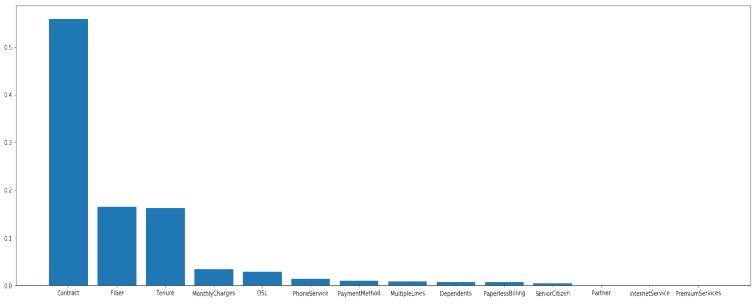

Figure 14: Features Importance model 1.

We can observe how the clusters found previously are found in the decision tree. In fact we find the class of those who do not have fiber but have telephone services, which tend not to leave behind a low threshold of tenure (Cluster 0 with respect to DBSCAN labeling). We find in the central and more articulated part of the tree (Cluster 1), the large part of the population of the 'Churn', characterized by subscription to fiber and premium services and with a short-

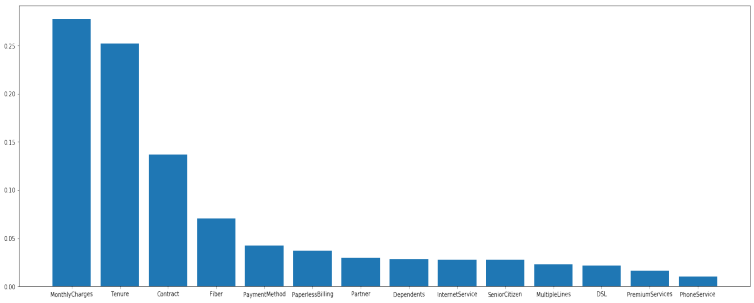

Figure 15: Features Importance model 2.

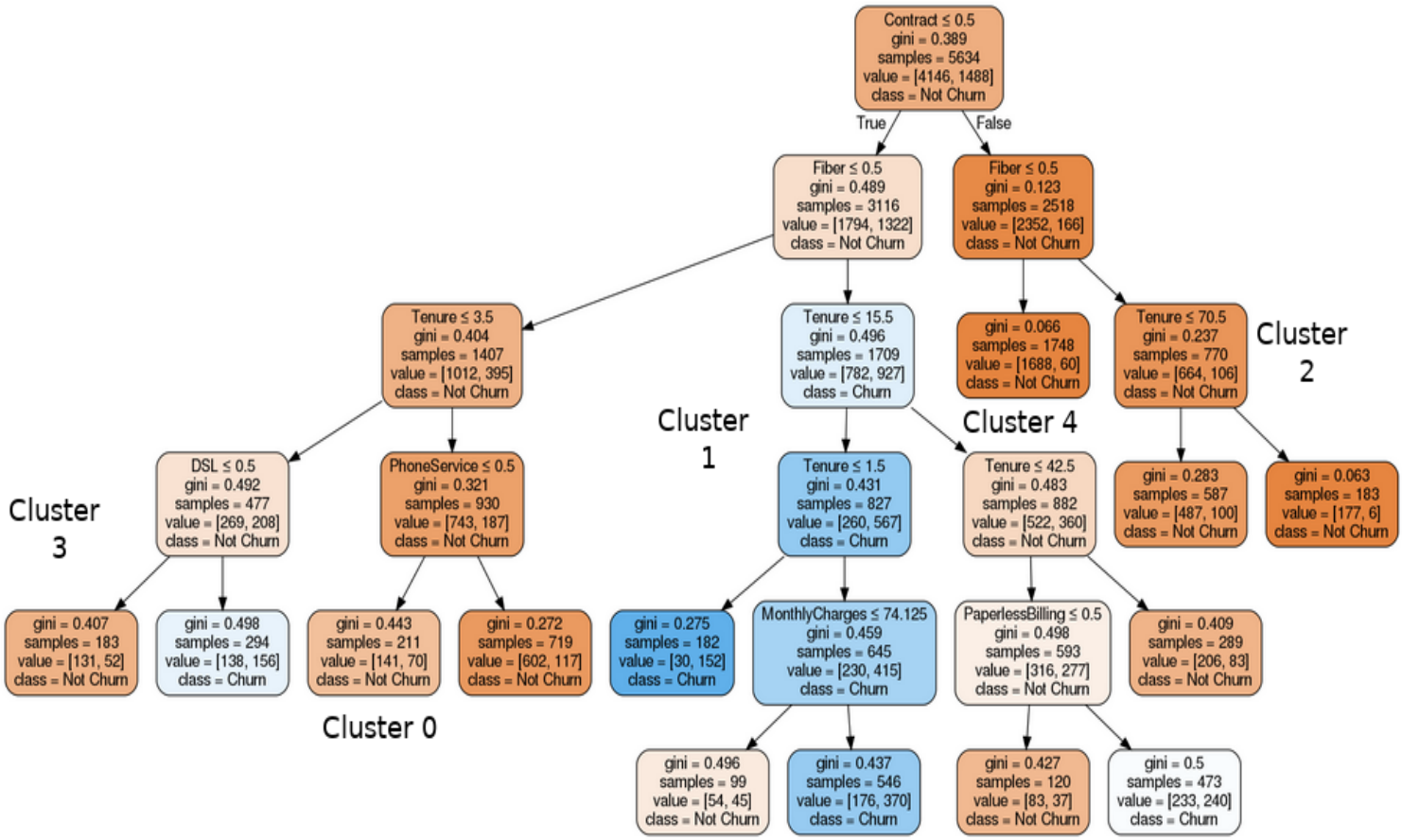

Figure 16: Decision Tree.

term contract. In the right branch of the decision tree there are those who have a long-term contract, fiber optic internet and a high Monthly Charge. These are the loyal customers (Cluster 2). Finally we find the last two similar groups of 'not churn' customers, those that have opted for DSL internet connection type and short-term contract (Cluster 3) and those that are long-term clients with high monthly charge contracts (Cluster 4).

5 Conclusion

To conclude, we could suggest the company to focus on new-customer with high-end contracts by improving their internet service or also by lowering their fee in order to retain an even higher number of clients. Pertaining the dataset, a more balanced dataset would have resulted in better results. To reduce its skewness towards non-churning customers we could have also applied over-sampling techniques in order to generate new synthetic churning data points. In classification task, we have also tried to build a fine-tailored training set by including in it an higher density of churning customer but without significant improvements in results.