[IntroToML] First Project - report

Marco Cocciaretto

May 2025

1 Introduction

This project encompasses many areas of machine learning and focuses on finding models able to fit the dataset Telco Customer Churn.

2 Data exploration and preparation

The first part of the project (and, really, of any data analysis work) consists in understanding the data that we are dealing with, fixing the issues that we may recognize with it, visualizing it and manipulating it to better fit our needs. In the case of our dataset, this process amounted to:

- Checking for missing values,
- Convert categorical values to integer or boolean,
- Normalize numerical data,
- Explore distributions and correlations,
- Create new features and drop useless ones.

2.1 Check for missing values

Luckily, our dataset is relatively clean; actually, when blindly running `telco_df.info()` to check for missing values, we might be mislead into thinking that there aren't any. However, looking at the `TotalCharges` feature, something feels off: we expect it to be a `float64`, instead it displays as an `object` type. It turns out that some entries have an empty space ' ' under that feature, possibly due to the customer not having paid their fee yet (this is further exacerbated by the fact that, for those entries, the corresponding `tenure` is 0). Therefore, we replace those values with 0.

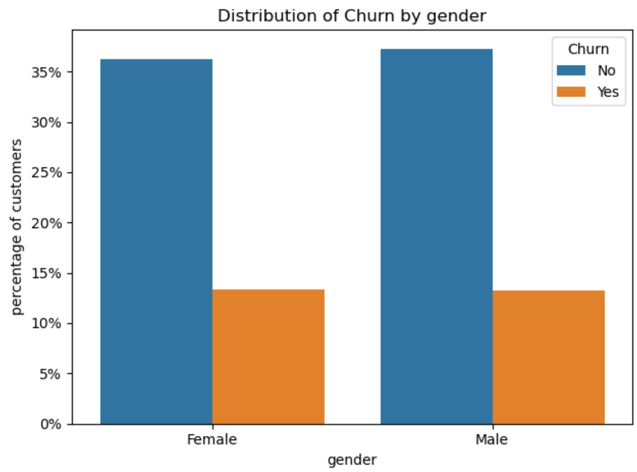

Figure 1: Distribution of churn relative to customer gender

2.2 Convert categorical values into integers

This step is mostly self-explanatory. There is, however, some nuance in how it was implemented: multiple copies of the dataset where imported and different dataframes corresponded to different data preparation processes. For example, on the "standard" dataframe `telco_df` the categorical-to-numerical mapping was performed by assigning a numerical value for each categorical value, while for the dataframe `telco_df_ar`, used for association rule mining (Section 5), we used One-Hot encoding.

2.3 Normalize data

Numerical data is often defined on wildly different ranges, which may hurt the performance of the model that gets trained on it. For this reason, it's customary to "normalize" this data, effectively rescaling its values. The `sklearn` built-in `StandardScaler` was chosen for this task.

2.4 Explore distributions and correlations

Understanding how the data is distributed and its correlations may help reveal important features and useless ones. For example, by plotting the churn distribution relative to gender (see Figure 1), we can see that the feature clearly does not impact customer churn, as we would expect. We can also plot a correlation heatmap to visualize compactly feature correlations, as done in Figure 2.

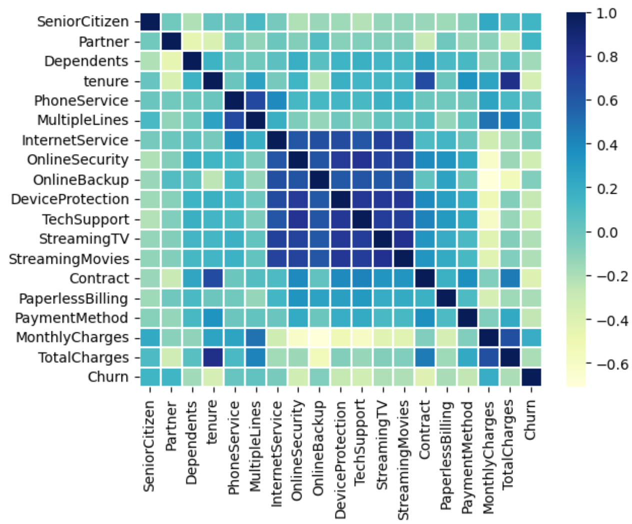

Figure 2: Correlation heatmap

2.5 Feature dropping and creation

Useless features (e.g. `gender`, `CustomerID`) were dropped during this step. Some new features, i.e. a binning of `tenure` and the difference:

$$\texttt{TotalCharges} - \texttt{MonthlyCharges} * \texttt{tenure} \tag{1}$$

were added to some of the copies of the dataset. For one of those copies, we only retained the 10 most important features based on a preliminary `Random Forest` training.

3 Unsupervised learning

We ran the following clustering algorithms:

- Agglomerative clustering, both with Ward and single linkage
- K-means
- DBSCAN

For each of these, we used a knee method estimation to find an OK guess for the model parameters (cluster number or `eps`, an example is reported in Figure 3). None of these methods yielded strong evidence of the presence of natural clusterings, we summarize the reasons why in the list below:

- Internal evaluation: the internal metrics used for the analysis where the SSE (used mainly for knee method estimation), the silhouette and the

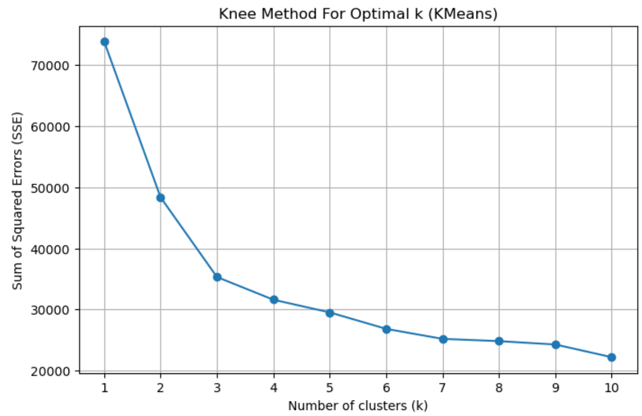

Figure 3: Knee method estimation for the number of clusters in a K-means algorithm, in which we need to look for the "bending" point, in this case 3.

separation. The best combined scores we obtained for 3 clusters were *silhouette* ~ 0.32 (decent but not great) and *separation* ~ 1.27, which is far from < 1 score we hoped for. We obtained better results for 2 clusters only for agglomerative clustering, reaching *silhouette* ~ 0.38 and *separation* ~ 0.95; this clustering, however, did not correspond to the target variable, as the external scores were not good at all.

- External evaluation: here we selected the metrics ARI, homogeneity, completeness and V_measure. For all of these, we have never obtained a scoring above ~ 0.17, meaning that the clustering does not predict the target variable.

Overall, these scores indicate some clustering tendency but with a lot of overlap. This could have been predicted: customers of internet and telephone services providers most likely have overlapping needs and wants, which would make a clustering algorithm not ideal; furthermore, if we were to only think about the target variable (even though in this context we should forget about it) the reasons behind churning can vary and likely overlap with each other, as can we can also see in the results reported in Section 5.

4 Supervised learning

Here we obtained some decent results, which are summarized in Figure 4. More detailed metrics are reported in the notebook.

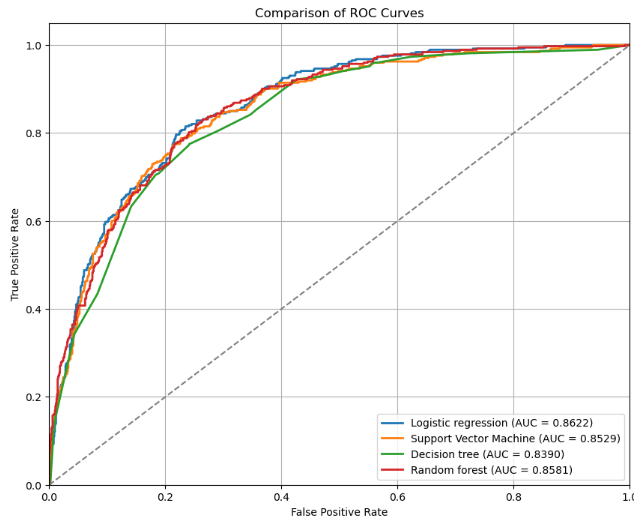

Figure 4: Comparison of ROC for different models trained on the dataframes; the performances of a given model trained on different dataframes were very similar, even on the lighter one where only 10 feature were not dropped.

Rule	**Support**	**Confidence**	**Lift**
OnlineSecurity_No, tenure_bin_Low ⇒ Churn_1, Contract_Month-to-month	0.107767	0.626755	2.667211
TechSupport_No, tenure_bin_Low ⇒ Churn_1, Contract_Month-to-month	0.106489	0.620347	2.639944
Contract_Month-to-month, InternetService_Fiber optic, OnlineSecurity_No, PaymentMethod_Electronic check ⇒ Churn_1, PhoneService	0.101661	0.634752	2.631287
Contract_Month-to-month, InternetService_Fiber optic, PaymentMethod_Electronic check, TechSupport_No ⇒ Churn_1, PhoneService	0.101661	0.629174	2.608165
Contract_Month-to-month, InternetService_Fiber optic, OnlineBackup_No, OnlineSecurity_No ⇒ Churn_1, PhoneService	0.106489	0.610252	2.529727
Contract_Month-to-month, InternetService_Fiber optic, OnlineBackup_No, TechSupport_No ⇒ Churn_1, PhoneService	0.107199	0.609855	2.528079
Contract_Month-to-month, InternetService_Fiber optic, OnlineSecurity_No, TechSupport_No ⇒ Churn_1, PhoneService	0.131336	0.606955	2.516060
Contract_Month-to-month, DeviceProtection_No, InternetService_Fiber optic, OnlineSecurity_No ⇒ Churn_1, PhoneService	0.104359	0.603944	2.503578
Contract_Month-to-month, InternetService_Fiber optic, PaymentMethod_Electronic check ⇒ Churn_1, PhoneService	0.112026	0.603673	2.502452
Contract_Month-to-month, InternetService_Fiber optic, OnlineSecurity_No, PaymentMethod_Electronic check ⇒ Churn_1	0.101661	0.634752	2.391951

Table 1: Association Rules with positive churn in the consequents

Rule	**Support**	**Confidence**	**Lift**
Churn_0, DeviceProtection_No internet service, InternetService_No, MultipleLines_No, OnlineSecurity_No internet service, PhoneService, StreamingTV_No internet service, TechSupport_No internet service ⇒ OnlineBackup_No internet service, PaperlessBilling, SeniorCitizen_0	0.11018	0.717854	4.833503
Churn_0, DeviceProtection_No internet service, InternetService_No, MultipleLines_No, OnlineBackup_No internet service, OnlineSecurity_No internet service, PhoneService, StreamingTV_No internet service ⇒ PaperlessBilling, SeniorCitizen_0, TechSupport_No internet service	0.11018	0.717854	4.833503
Churn_0, DeviceProtection_No internet service, MultipleLines_No ⇒ InternetService_No, MonthlyCharges_bin_Low, OnlineBackup_No internet service, OnlineSecurity_No internet service, PaperlessBilling, PhoneService, SeniorCitizen_0, StreamingMovies_No internet service, StreamingTV_No internet service, TechSupport_No internet service	0.11018	0.717854	4.833503
Churn_0, MonthlyCharges_bin_Low, MultipleLines_No ⇒ DeviceProtection_No internet service, InternetService_No, OnlineBackup_No internet service, OnlineSecurity_No internet service, PaperlessBilling, PhoneService, SeniorCitizen_0, StreamingMovies_No internet service, StreamingTV_No internet service, TechSupport_No internet service	0.11018	0.717854	4.833503
Churn_0, MultipleLines_No, StreamingTV_No internet service ⇒ DeviceProtection_No internet service, InternetService_No, MonthlyCharges_bin_Low, OnlineBackup_No internet service, OnlineSecurity_No internet service, PaperlessBilling, PhoneService, SeniorCitizen_0, StreamingMovies_No internet service, TechSupport_No internet service	0.11018	0.717854	4.833503
Churn_0, MultipleLines_No, TechSupport_No internet service ⇒ DeviceProtection_No internet service, InternetService_No, MonthlyCharges_bin_Low, OnlineBackup_No internet service, OnlineSecurity_No internet service, PaperlessBilling, PhoneService, SeniorCitizen_0, StreamingMovies_No internet service, StreamingTV_No internet service	0.11018	0.717854	4.833503
Churn_0, MultipleLines_No, OnlineBackup_No internet service ⇒ DeviceProtection_No internet service, InternetService_No, MonthlyCharges_bin_Low, OnlineSecurity_No internet service, PaperlessBilling, PhoneService, SeniorCitizen_0, StreamingMovies_No internet service, StreamingTV_No internet service, TechSupport_No internet service	0.11018	0.717854	4.833503
Churn_0, InternetService_No, MultipleLines_No ⇒ DeviceProtection_No internet service, MonthlyCharges_bin_Low, OnlineBackup_No internet service, OnlineSecurity_No internet service, PaperlessBilling, PhoneService, SeniorCitizen_0, StreamingMovies_No internet service, StreamingTV_No internet service, TechSupport_No internet service	0.11018	0.717854	4.833503
Churn_0, InternetService_No, MultipleLines_No, OnlineBackup_No internet service, OnlineSecurity_No internet service, PhoneService, StreamingTV_No internet service, TechSupport_No internet service ⇒ DeviceProtection_No internet service, PaperlessBilling, SeniorCitizen_0	0.11018	0.717854	4.833503
Churn_0, DeviceProtection_No internet service, InternetService_No, MultipleLines_No, OnlineBackup_No internet service, OnlineSecurity_No internet service, PhoneService, TechSupport_No internet service ⇒ PaperlessBilling, SeniorCitizen_0, StreamingTV_No internet service	0.11018	0.717854	4.833503

Table 2: Association Rules with no condition on the consequents

5 Association rule mining

The association rule mining was performed using the `apriori` algorithm for the frequent itemset generation. We report the best rules (based on lift) with no conditions on the consequents in Table 2 and the best rules with positive churn in the consequents in Table 1. As previously anticipated, we find rules with good metrics and different but overlapping antecedents which can reliably predict churn, possibly providing an explanation for the clustering failures. Interestingly, most rules' antecedents include the fact that the customer has an optic fiber internet service, telling us that most churning customers might be more tech savvy or have more demanding needs for their ISP's. Each of the best rules overall involve a negative churn in the antecedent, and the corresponding consequents have either or both the properties of low monthly charge and the customer not being a senior citizen.