Scientific Journal of Astana IT University
ISSN (P): 2707-9031
ISSN (E): 2707-904X

Bauyrzhan Mubarak
Big Data Analysis, Astana IT University
211508@astanait.edu.kz
Kazakhstan
Sekenov Gabit
Big Data Analysis, Astana IT University
201308@astanait.edu.kz
Kazakhstan

Customer Churn Prediction in Telecommunication

Abstract: The main task of customer churn prediction is to estimate subscribers who may want to leave from a company and provide solutions to prevent possible churnes. In recent years, estimating churners before they leave has become valuable in the environment of increased competition among companies. The research in this paper was done to estimate what clustering algorithms are appropriate for the dataset of churn prediction and metrics that show the level of implementation of clustering algorithms.

Keywords: Customer Churn Prediction; Clustering algorithms; Machine Learning; Model Selection; Model training; Model evaluation

Introduction (Literary review)

Customer churn has become highly important for companies because of increasing competition among companies, increased importance of marketing strategies and conscious behavior of customers in recent years. Customers can easily trend toward alternative services. Companies must develop various strategies to prevent these possible trends, depending on the services they provide. During the estimation of possible churns, data from the previous churns might be used. An efficient churn predictive model benefits companies in many ways. Early identification of customers likely to leave may help to build cost effective ways in marketing strategies. Customer retention campaigns might be limited to selected customers but it should cover most of the customer. Incorrect predictions could result in a company losing profits because of the discounts offered to continuous subscribers. Therefore, the right predictions of the churn customers has become highly important for the companies.

A. What are clustering algorithms?

Clustering algorithms are a category of unsupervised machine learning techniques that aim to group similar data points into clusters or subgroups based on certain criteria. The goal of clustering is to identify inherent patterns or structures within the data without using predefined labels.

B. What is SVM?

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. SVM is particularly effective in high-dimensional spaces and is capable of handling both linear and non-linear relationships between input features and output classes.

The primary objective of SVM is to find a hyperplane that best separates data points belonging to different classes in a feature space. This hyperplane is chosen to maximize the margin, which is the distance between the hyperplane and the nearest data points from each class, also known as support vectors. Maximizing the margin helps improve the generalization ability of the model.

Data preprocessing and exploration:

Data preprocessing is a crucial step in the data analysis and machine learning pipeline. It involves cleaning and transforming raw data into a format that can be effectively used for analysis or training machine learning models. Proper data preprocessing can improve the performance and reliability of models. Here are some common steps in data preprocessing:

A. General data information:

```
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   customerID        7043 non-null   object
 1   gender            7043 non-null   object
 2   SeniorCitizen     7043 non-null   int64
 3   Partner           7043 non-null   object
 4   Dependents        7043 non-null   object
 5   tenure            7043 non-null   int64
 6   PhoneService      7043 non-null   object
 7   MultipleLines     7043 non-null   object
 8   InternetService   7043 non-null   object
 9   OnlineSecurity    7043 non-null   object
 10  OnlineBackup      7043 non-null   object
 11  DeviceProtection  7043 non-null   object
 12  TechSupport       7043 non-null   object
 13  StreamingTV       7043 non-null   object
 14  StreamingMovies   7043 non-null   object
 15  Contract          7043 non-null   object
 16  PaperlessBilling  7043 non-null   object
 17  PaymentMethod     7043 non-null   object
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object
 20  Churn             7043 non-null   object
```

B. Handling missing values:

Identify and handle missing data points. This can involve removing instances with missing values, imputing missing values (replacing them with estimated values), or using techniques like interpolation.

```
df[df.columns[df.isnull().any()]].isnull().sum()

TotalCharges    11
dtype: int64
```

Handling missing values by the mode is a common technique, especially for categorical data where the mode is the most frequently occurring value. The mode represents the central tendency for categorical variables, and replacing missing values with the mode can be a reasonable strategy.

```
mode = df.TotalCharges.mode()[0]
mode

20.2

df['TotalCharges'] = df['TotalCharges'].replace(np.nan,mode) #Пропущенные заполняю модой

df[df.columns[df.isnull().any()]].isnull().sum()

Series([], dtype: float64)
```

C. *Feature engineering:*

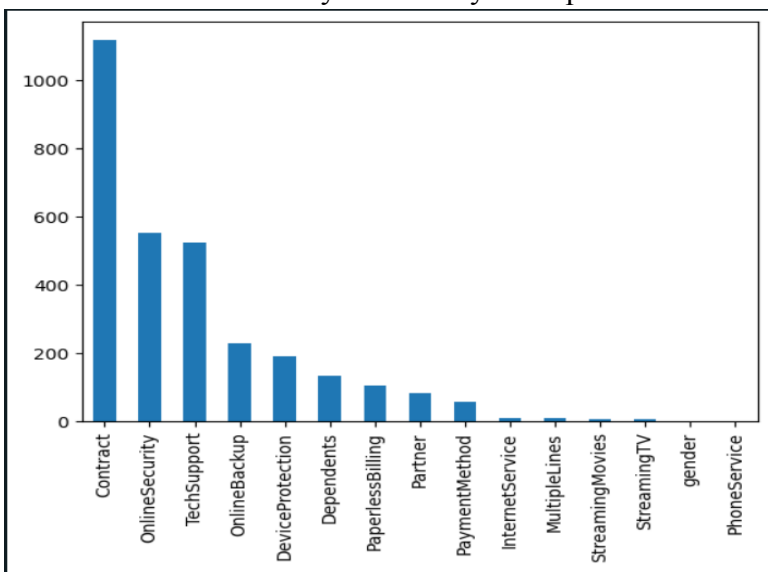
The chi-square (χ^2) test is a statistical test used to determine whether there is a significant association between two categorical variables. In the context of feature selection, the chi-square test can be used to assess the independence between each feature and the target variable in a classification problem. Features that are independent of the target variable may be considered less relevant for the task.

Calculate the chi-square statistic using this formula:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where O is the observed frequency, and E is the expected frequency.

Visualization of necessary features by its importance:



D. *Feature Scaling:*

Standardize or normalize numerical features to bring them to a similar scale. This is important for algorithms sensitive to the magnitude of features, such as gradient descent-based optimization algorithms.

```
#Нормализируем данные
df['TotalCharges'] = df['TotalCharges']/df['TotalCharges'].max()
df['tenure'] = df['tenure']/df['tenure'].max()
df['MonthlyCharges'] = df['MonthlyCharges']/df['MonthlyCharges'].max()
```

✓ 0.0s

E. *Changing data types to appropriate ones:*

Some features have some problems with data types as object types which should be floats.

```
def transfomFloat(x):
    try:
        return float(x)
    except:
        return np.nan
df['TotalCharges']= df['TotalCharges'].apply(transfomFloat) #Column TotalCharges менять на float
```

✓ 0.0s

Model selection:

K-Means is a popular clustering algorithm that partitions data into k clusters. The algorithm aims to minimize the variance within each cluster by iteratively updating the cluster centroids and assigning data points to the nearest centroid.

- It is a centroid-based algorithm.
- Partitions the data into k clusters based on mean points.
- Requires specifying the number of clusters (k) in advance.

The hierarchical agglomerative clustering uses the bottom-up approaches. The HAC algorithm starts with every single data point as a single cluster. The similar clusters are successively merged until all clusters have merged into a one cluster and the result is represented in tree structure as named dendrogram.

- Builds a tree of clusters by iteratively merging or splitting existing clusters.
- Provides a hierarchy of clusters, which can be visualized as a dendrogram.

Density-Based Spatial Clustering of Applications with Noise, commonly referred to as DBSCAN, is a clustering algorithm used in unsupervised learning. Its primary function is to group together data points that are densely packed, meaning they have many nearby neighbors. This algorithm is particularly useful in identifying outliers within a data set, marking them as noise.

- It is a density-based algorithm.
- Clusters data based on density, identifying regions with higher point density.
- Can discover clusters of arbitrary shapes and handles noise well.

A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. One can think of mixture models as generalizing k-means clustering to incorporate information about the covariance structure of the data as well as the centers of the latent Gaussians.

- It is a probabilistic model.
- Assumes that the data is generated from a mixture of several Gaussian distributions.
- Each data point belongs to a cluster with a certain probability.

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is a clustering algorithm that can cluster large datasets by first generating a small and compact summary of the large dataset that retains as much information as possible. This smaller summary is then clustered instead of clustering the larger dataset.

- It is a hierarchical clustering method.
- Designed for large datasets with an online, memory-efficient approach.
- Builds a tree structure to represent the data distribution.

Results:
Performance of clustering models were measured through silhouette score:

Silhouette Score is a metric to evaluate the performance of a clustering algorithm. It uses compactness of individual clusters(***intra cluster distance***) and separation amongst clusters (***inter cluster distance***) to measure an overall representative score of how well our clustering algorithm has performed.This is a simple metric but many a times we fail to use it correctly ending up quoting numbers which are false representation of actual score. In this blog I will cover a simple example of how silhouette score may be misleading if used blindly. To keep things visually understandable I will stick to two dimensional dataset but the idea can be promoted for multidimensional dataset as well.

Silhouette Score for a datapoint **i** is given as

$$s_i = \frac{b_i - a_i}{max(b_i, a_i)}$$

where,
bi : is the inter cluster distance defined as the average distance to closest cluster of datapoint i except for that it's a part of

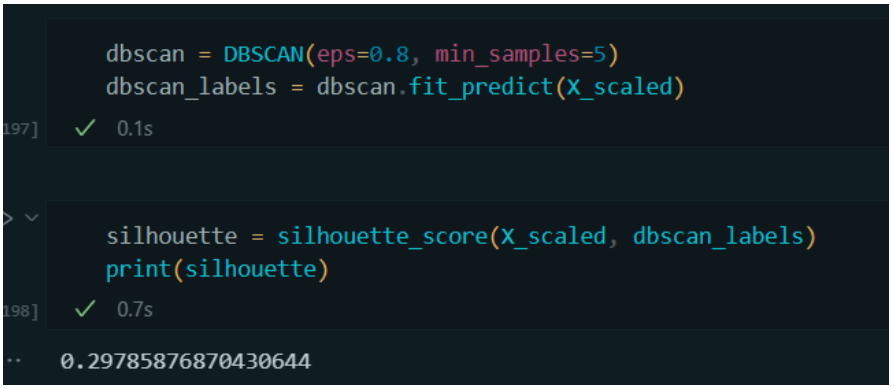
$$b_i = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)$$

ai : is the intra cluster distance defined as the average distance to all other points in the cluster to which it's a part of

$$a_i = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

ai : is the intra cluster distance defined as the average distance to all other points in the cluster to which it's a part of

In general,from 5 algorithms 3 algorithms had the same silhouette score as K-Means,Agglomerative Clustering and Birch.DBSCAN has the highest score among them,having nearly 30%.OPTICS algorithm has the lowest score,reaching nearly 10%.



Conclusion:
In conclusion, this study demonstrates the means and methods of creating clustering models with the results of such a model using a customer data set as an example and churn prediction as a target. Additionally this paper provided descriptions of various clustering models, namely K-clustering, agglomerate, DBSCAN, GMM and BIRCH. Overall, the results show the potential of clustering algorithms as well as differences in comparison to other models. Not to mention the fact that such models are highly more adaptive than regular ones and can be easily adapted depending on the user's needs showing that such models are a future step for creating and developing ML models.

References:

- *Zhang T., Ramakrishnan R., Livny M.* BIRCH: an efficient data clustering method for very large databases // Proceedings of the 1996 ACM SIGMOD international conference on Management of data - SIGMOD '96. — 1996. — doi:10.1145/233269.233324.
- Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M. (eds.). *A density-based algorithm for discovering clusters in large spatial databases with noise* (PDF). Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231. CiteSeerX 10.1.1.121.9220. ISBN 1-57735-004-9.
- Kriegel, Hans-Peter; Schubert, Erich; Zimek, Arthur (2016). "The (black) art of runtime evaluation: Are we comparing algorithms or implementations?". *Knowledge and Information Systems*. **52** (2): 341–378. doi:10.1007/s10115-016-1004-2. ISSN 0219-1377. S2CID 40772241.
- MacQueen, J. B. (1967). *Some Methods for classification and Analysis of Multivariate Observations*. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Vol. 1. University of California Press. pp. 281–297. MR 0214227. Zbl 0214.46201. Retrieved 2009-04-07.
- Cortes, Corinna; Vapnik, Vladimir (1995). "Support-vector networks" (PDF). *Machine Learning*. **20** (3): 273–297. CiteSeerX 10.1.1.15.9362. doi:10.1007/BF00994018. S2CID 206787478.
- Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2008). *The Elements of Statistical Learning : Data Mining, Inference, and Prediction* (PDF) (Second ed.). New York: Springer. p. 134.