CHURN PREDICTION IN TELECOM INDUSTRY

Darya Solomenko

NCI Machine Learning - June 2024

1. Introduction

Business profitability relies heavily on customer loyalty, which can be difficult to achieve in the telecommunications industry. Customer attrition (also known as churn) has become the most common concern in the telecom business, due to a more competitive market and the possibility for customers to transfer providers in minutes.

Churn rate refers to the percentage of subscribers who cancel their contract or subscription with a telecom company within a specified timeframe [1]. A high churn rate implies a failure in customer retention, reflecting a lack of loyalty to firm’s product or service.

If a Telecom company wishes to survive in the competitive market customer retention should become their long-term marketing objective and overall company goal [2]. Studies shown that by focusing on customer loyalty Telecom business owners can not only retain clients but also increase profits [3]. Customer relationship management (CRM) is referred by Bill Lee as “the greatest overlooked opportunity” for the business growth [4]. While CRM holds a huge potential for Telecom industry churn prediction is one of the ways to better understand customer behaviour patterns and identify customers at risk.

The most common reasons of churn in the telecom industry are poor service experience, negative customer experience, unclear billing and better deals from the competitors [5]. While it is useful to understand deeper reasons of clients’ attrition, it is also worthwhile to deploy machine learning (ML) algorithms to predict churn. Studies show that accurate churn prediction can be achieved using various ML classifier models like decision tree and neural network [6]. ML model’s output can be used for personalized interventions by customer support team or promotions (offers or discounts) which may lead to higher customer satisfaction and as a result churn reduction.

The goals of this project are to:

- Perform an exploratory data analysis of a Telecom customers dataset and verify that the dataset is suitable for ML application.
- Pre-process data in preparation for ML model fitting.
- Select and fit a suitable classifier ML model on the pre-processed data and evaluate model performance.
- Assess ethical considerations related to data used, pre-processing approach and model application.

2. Dataset

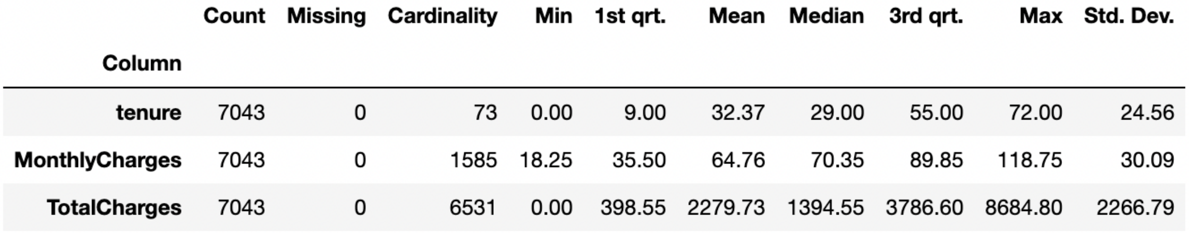
The dataset used for analysis - Telco Customer Churn [7] by IBM - includes information about an imaginary telecommunication company that provided home phone and Internet services to in California in Q3 2018. The dataset contains records for 7,043 customers and 21 variables detailed in the Appendix 1.

Exploratory Data Analysis

2-1. Identifiers: ‘**customerID**’: 7,043 unique string values.

2-2. **Continuous features:** there are three continuous features in the dataset: ‘tenure’, ‘MonthlyCharges’, and ‘TotalCharges’. Their descriptive statistics is shown in Table 1 and visualised in **Error! Reference source not found.**.

Table 1 Key characteristics of continuous features



Column	Count	Missing	Cardinality	Min	1st qrt.	Mean	Median	3rd qrt.	Max	Std. Dev.
tenure	7043	0	73	0.00	9.00	32.37	29.00	55.00	72.00	24.56
MonthlyCharges	7043	0	1585	18.25	35.50	64.76	70.35	89.85	118.75	30.09
TotalCharges	7043	0	6531	0.00	398.55	2279.73	1394.55	3786.60	8684.80	2266.79

All three numeric features seem to have non normal distribution based on the shape of the histogram. There were no significant outliers among the values for the three features and their normalised variance was comparable (< 1). As data ranges for ‘tenure’, ‘MonthlyCharges’, and ‘TotalCharges’ vary largely, scaling needs to be performed on pre-processing stage for compensate for it.

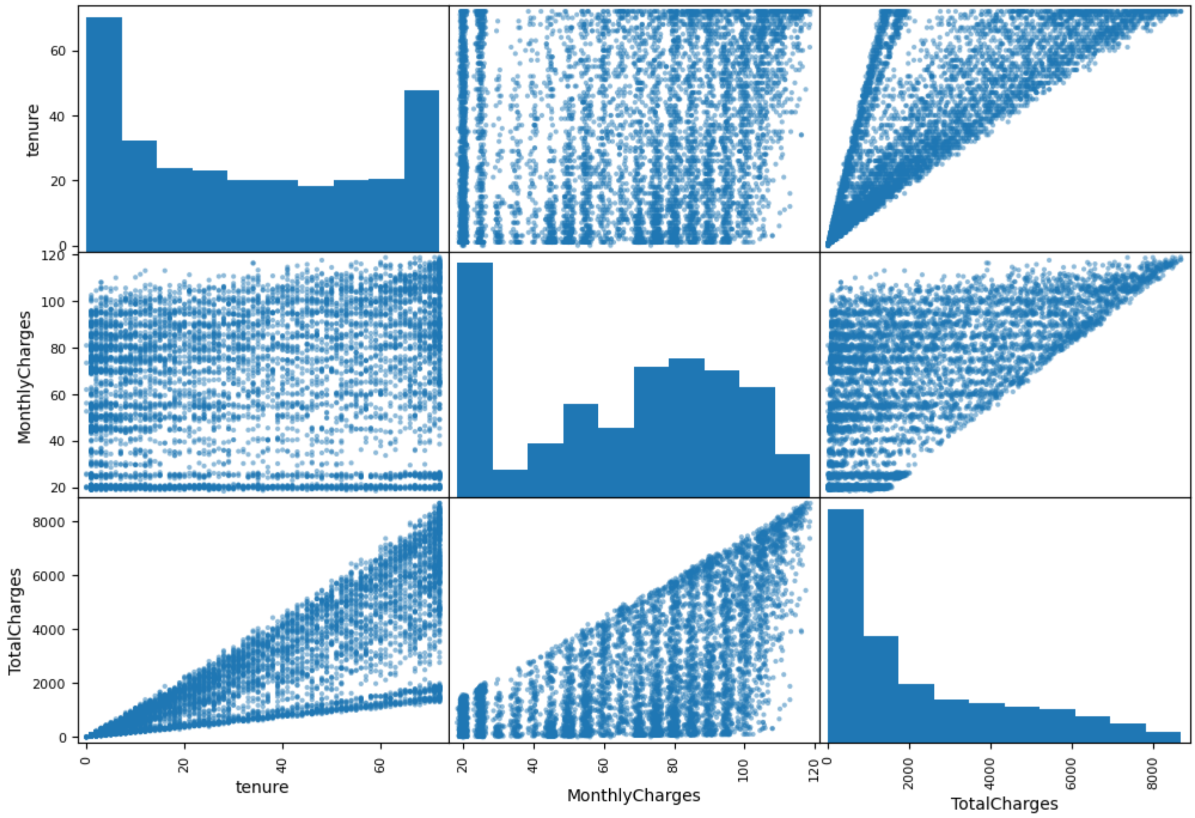


Figure 1 Histograms and scatter plots for continuous features

Scatterplots (Figure 1) suggest that there might be moderate to strong correlation between ‘TotalCharges’ and two other numeric features. Correlation matrix (Figure 2) confirms this assumption: Pearson coefficient for ‘TotalCharges’ and ‘tenure’ is 0.83 (strong positive correlation) and for ‘TotalCharges’ is 0.65 (moderate positive correlation). Absence of multicollinearity is one of common assumptions for ML models and need to be considered during feature selection process.

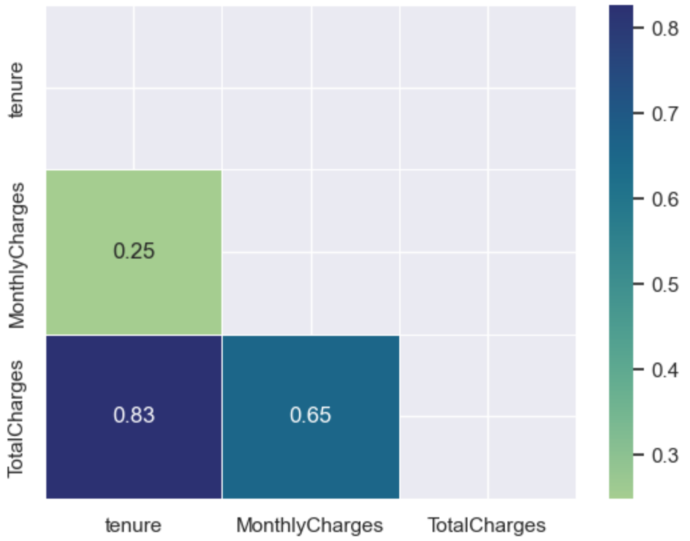


Figure 2 Correlation matrix for numeric features

2-3. Categorical features:

2-3.1. Binary:

2-3.1.1. **"Gender":** Male 3,555 (50.5%); Female 3,488 (49.5%) – nearly equal split.

2-3.1.2. **"Churn":** Yes 1,869 (26.5%); No 5,174 (73.5%) – around a quarter of customers in dataset churned. As churn is a predictor value it shows that dataset is unbalanced which needs to be considered when fitting and evaluating the model.

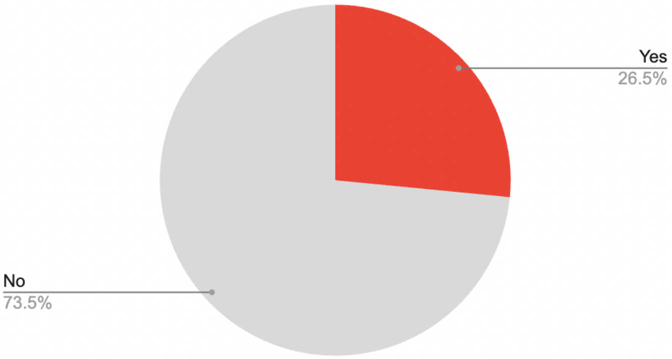


Figure 3 Split of records by predictor (churn)

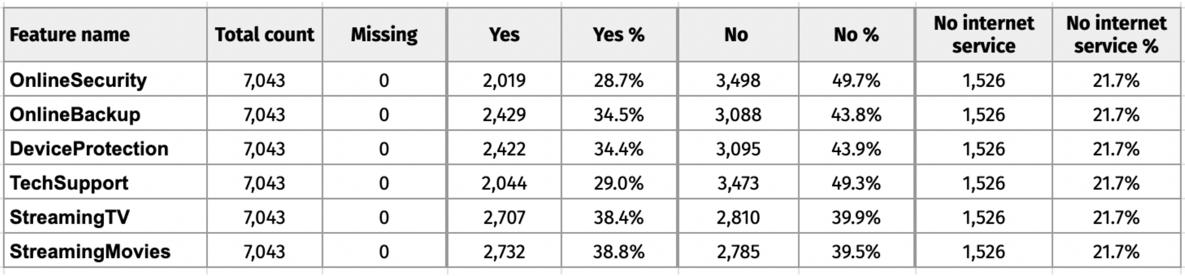
2-3.2. With three or more options:

2-3.2.1. **"MultipleLines":** Yes 2,971 (42.2%); No 3,390 (48.1%), No phone service 682 (9.7%) – majority of customers have phone service with single line.

2-3.2.2. **"InternetService":** DSL 2,421 (34.4%); Fiber Optic 3,096 (44.0%); No 1,526 (21.7%) – slightly more than ¼ of customers do not have internet service. Fiber Optic is more popular internet option than DSL.

2-3.2.3. **Internet services:** additional services used by customers with internet services are listed in Table 2.

Table 2 Count and share of categories for internet services



Feature name	Total count	Missing	Yes	Yes %	No	No %	No internet service	No internet service %
OnlineSecurity	7,043	0	2,019	28.7%	3,498	49.7%	1,526	21.7%
OnlineBackup	7,043	0	2,429	34.5%	3,088	43.8%	1,526	21.7%
DeviceProtection	7,043	0	2,422	34.4%	3,095	43.9%	1,526	21.7%
TechSupport	7,043	0	2,044	29.0%	3,473	49.3%	1,526	21.7%
StreamingTV	7,043	0	2,707	38.4%	2,810	39.9%	1,526	21.7%
StreamingMovies	7,043	0	2,732	38.8%	2,785	39.5%	1,526	21.7%

While typically there are more customers that have not sign up for additional internet services, opt-in share (yes %) varies between 28.7% and 38.8% which indicates different demand for the services.

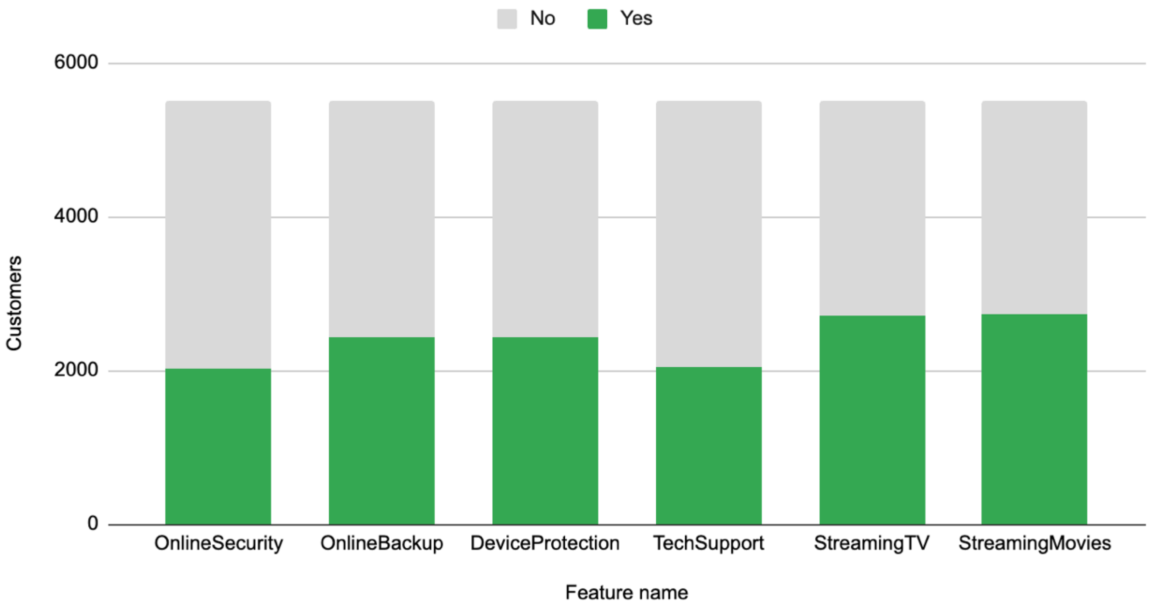


Figure 4 Feature adoption among customers with Internet service

2-3.2.4. **“Contract”:** Month-to-Month 3,875 (55.0%); One Year 1,473 (20.9%); Two Year 1,695 (24.1%) – over the half of the customers have month-to-month contract, with two-year contract being the second most popular category.

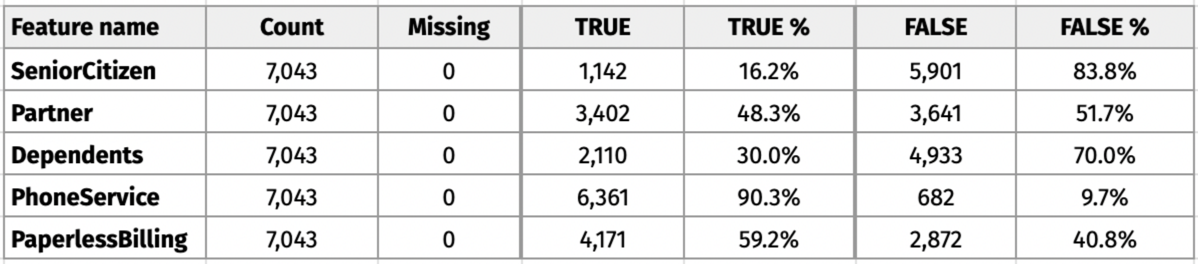
2-3.2.5. **“PaymentMethod”**: Electronic Check 2,365 (33.6%); Mailed Check 1,612 (22.9%); Credit Card (automatic) 1,522 (21.6%); Bank transfer (automatic) 1,544 (21.9%) – a third of customers are paying using Electronic Check with the rest of the categories nearly evenly split.

2-4. **Boolean features** – details on Boolean features in the dataset are listed in Table 3.

Boolean features describe both demographic characteristics (e.g. age group) of customers and service choices (e.g. paperless billing):

- Majority of customers are younger than 65 years (83.8%) and have no dependents (70%). There is nearly 50/50 split for presence of a partner.
- Vast majority (90.3%) of customers have phone service and 59.2% opted in for paperless billing.

Table 3 Count and share for Boolean features



Feature name	Count	Missing	TRUE	TRUE %	FALSE	FALSE %
SeniorCitizen	7,043	0	1,142	16.2%	5,901	83.8%
Partner	7,043	0	3,402	48.3%	3,641	51.7%
Dependents	7,043	0	2,110	30.0%	4,933	70.0%
PhoneService	7,043	0	6,361	90.3%	682	9.7%
PaperlessBilling	7,043	0	4,171	59.2%	2,872	40.8%

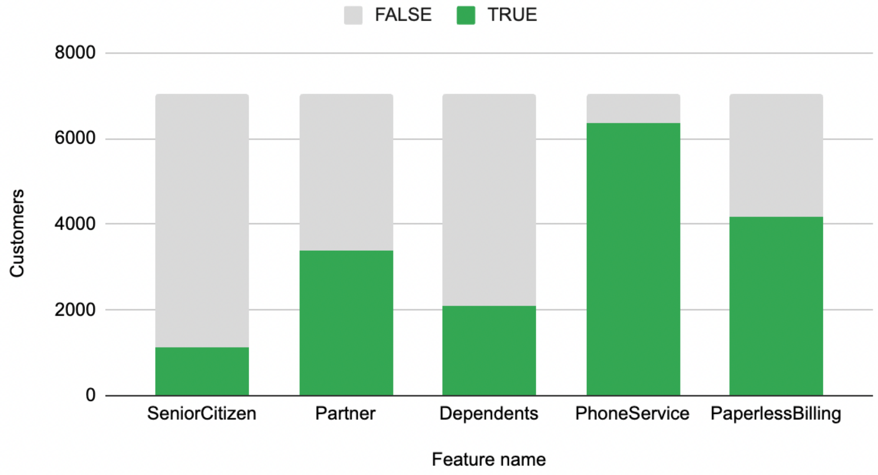


Figure 5 Distribution of values for Boolean features

3. Pre-processing

Exploratory data analysis shows that there are multiple categories of features in the dataset. This data needs to be appropriately prepared for use in a ML model.

While there was no missing data on the first observation, some of the columns were loaded in incorrect format. In particular, “TotalCharges” feature contained string values versus float. Conversion into the correct data type uncovered missing data for customers with tenure equal 0.

The following pre-processing steps were applied to prepare the dataset for ML model fitting:

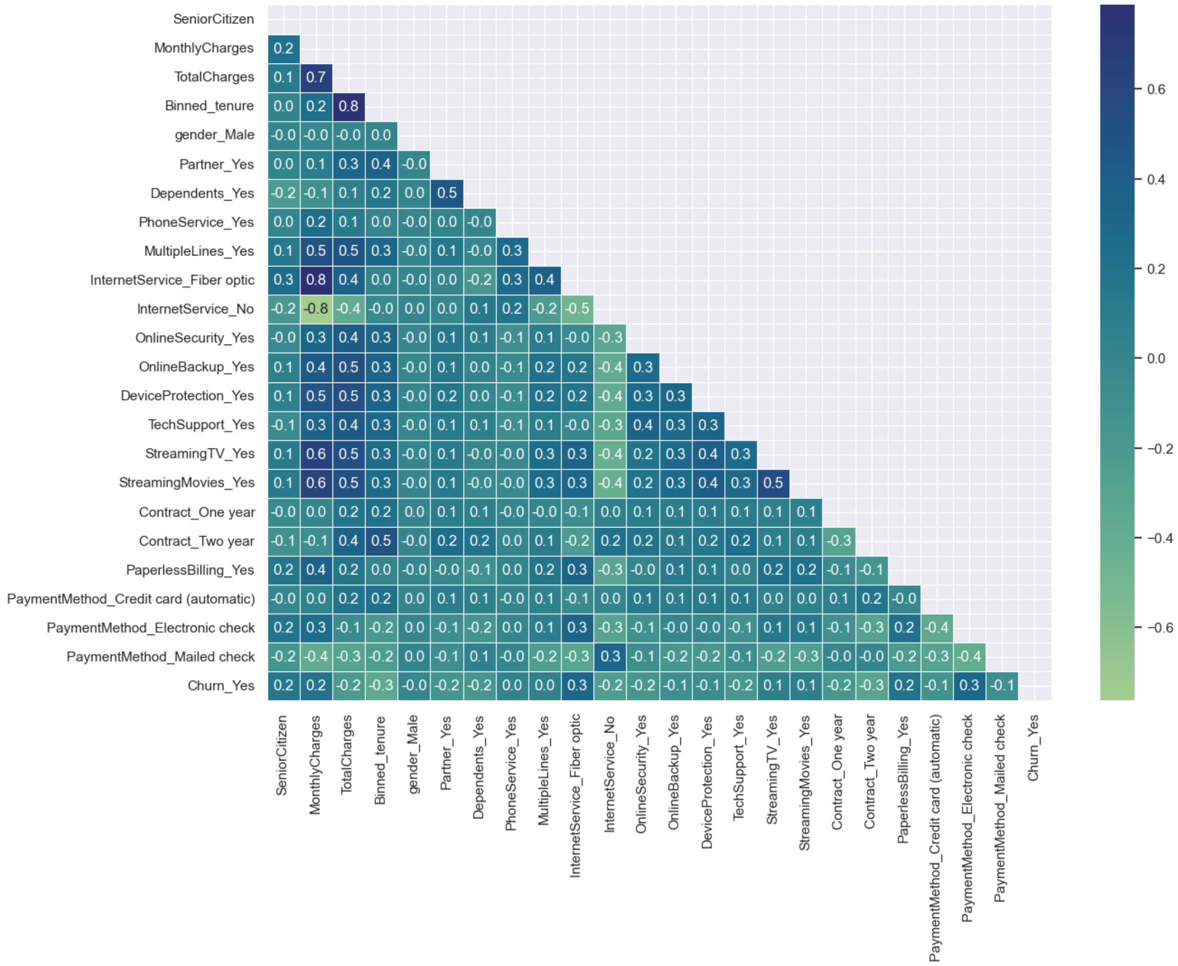
- **“customerID” feature dropped** as it was not useful for the model.
- **“TotalCharges” feature**:
  - Empty strings (‘ ‘) corresponding to customers with 0 tenure (11 rows) were replaced with 0 values for total charges.
  - The feature was converted to float format.
- New binned feature **“Binned_tenure”**[1] was created based on “tenure” with three groups: 1. under a year; 2. 1-4 years; 3. Over 4 years. As “tenure” data was not normally distributed with spikes at min and max, binning was performed for noise reduction.

[1] When tested on the model original “tenure” feature was preferred to the binned one as it resulted in better model performance.

- **All categorical and Boolean features** were replaced with dummy values using one-hot encoding method. First category dropped for each column. After encoding duplicating features were removed.
- Scaling transformation was applied to **continuous columns** as their ranges were vastly different which can affect model performance:
    - MinMaxScaler was preferred for numeric columns as it keeps the original distribution shape while scaling range down to between 0 and 1.
    - Scaler was applied after splitting data into train and text sets firstly to train data and then to test data. This was done to prevent data leakage and incorrect assessment of the model performance.

Correlation matrix for the resulting pre-processed dataset (24 features, 7043 rows) is shown in Figure 6. Some features show moderate to strong correlation (higher than 0.6 on less than -0.6). For example, “MonthlyCharges” positively correlates with “TotalCharges” (0.7), “InternetService_Fiber Optic” (0.8), “StreamingTV_Yes” (0.6), “StreamingMovies_Yes” (0.6) and negatively correlates with “InternetService_No” (-0.8). This means there is a potential either for feature selection (excluding some of the features from the model) or application of dimensionality reduction like Principal Component Analysis (PCA) to reduce number of features as there is a risk of overfitting the model.

It is suggested to try fitting model with data as is and after feature selection and/or dimensional reduction and compare model performance and select the best approach.



	SeniorCitizen	MonthlyCharges	TotalCharges	Binned_tenure	gender_Male	Partner_Yes	Dependents_Yes	PhoneService_Yes	MultipleLines_Yes	InternetService_Fiber optic	InternetService_No	OnlineSecurity_Yes	OnlineBackup_Yes	DeviceProtection_Yes	TechSupport_Yes	StreamingTV_Yes	StreamingMovies_Yes	Contract_One year	Contract_Two year	PaperlessBilling_Yes	PaymentMethod_Credit card (automatic)	PaymentMethod_Electronic check	PaymentMethod_Mailed check	Churn_Yes
SeniorCitizen																								
MonthlyCharges	0.2																							
TotalCharges	0.1	0.7																						
Binned_tenure	0.0	0.2	0.8																					
gender_Male	-0.0	-0.0	-0.0	0.0																				
Partner_Yes	0.0	0.1	0.3	0.4	-0.0																			
Dependents_Yes	-0.2	-0.1	0.1	0.2	0.0	0.5																		
PhoneService_Yes	0.0	0.2	0.1	0.0	-0.0	0.0	-0.0																	
MultipleLines_Yes	0.1	0.5	0.5	0.3	-0.0	0.1	-0.0	0.3																
InternetService_Fiber optic	0.3	0.8	0.4	0.0	-0.0	0.0	-0.2	0.3	0.4															
InternetService_No	-0.2	-0.8	-0.4	-0.0	0.0	0.0	0.1	0.2	-0.2	-0.5														
OnlineSecurity_Yes	-0.0	0.3	0.4	0.3	-0.0	0.1	0.1	-0.1	0.1	-0.0	-0.3													
OnlineBackup_Yes	0.1	0.4	0.5	0.3	-0.0	0.1	0.0	-0.1	0.2	0.2	-0.4	0.3												
DeviceProtection_Yes	0.1	0.5	0.5	0.3	-0.0	0.2	0.0	-0.1	0.2	0.2	-0.4	0.3	0.3											
TechSupport_Yes	-0.1	0.3	0.4	0.3	-0.0	0.1	0.1	-0.1	0.1	-0.0	-0.3	0.4	0.3	0.3										
StreamingTV_Yes	0.1	0.6	0.5	0.3	-0.0	0.1	-0.0	-0.0	0.3	0.3	-0.4	0.2	0.3	0.4	0.3									
StreamingMovies_Yes	0.1	0.6	0.5	0.3	-0.0	0.1	-0.0	-0.0	0.3	0.3	-0.4	0.2	0.3	0.4	0.3	0.5								
Contract_One year	-0.0	0.0	0.2	0.2	0.0	0.1	0.1	-0.0	-0.0	-0.1	0.0	0.1	0.1	0.1	0.1	0.1	0.1							
Contract_Two year	-0.1	-0.1	0.4	0.5	-0.0	0.2	0.2	0.0	0.1	-0.2	0.2	0.2	0.1	0.2	0.2	0.1	0.1	-0.3						
PaperlessBilling_Yes	0.2	0.4	0.2	0.0	-0.0	-0.0	-0.1	0.0	0.2	0.3	-0.3	-0.0	0.1	0.1	0.0	0.2	0.2	-0.1	-0.1					
PaymentMethod_Credit card (automatic)	-0.0	0.0	0.2	0.2	0.0	0.1	0.1	-0.0	0.1	-0.1	0.0	0.1	0.1	0.1	0.1	0.0	0.0	0.1	0.2	-0.0				
PaymentMethod_Electronic check	0.2	0.3	-0.1	-0.2	0.0	-0.1	-0.2	0.0	0.1	0.3	-0.3	-0.1	-0.0	-0.0	-0.1	0.1	0.1	-0.1	-0.3	0.2	-0.4			
PaymentMethod_Mailed check	-0.2	-0.4	-0.3	-0.2	0.0	-0.1	0.1	-0.0	-0.2	-0.3	0.3	-0.1	-0.2	-0.2	-0.1	-0.2	-0.3	-0.0	-0.0	-0.2	-0.3	-0.4		
Churn_Yes	0.2	0.2	-0.2	-0.3	-0.0	-0.2	-0.2	0.0	0.0	0.3	-0.2	-0.2	-0.1	-0.1	-0.2	0.1	0.1	-0.2	-0.3	0.2	-0.1	0.3	-0.1	

Figure 6 Correlation matrix for the pre-processed dataset

4. Machine Learning model

As churn prediction is a classification problem and the dependent variable ('Churn') is binary, logistic regression would be an appropriate ML model to apply.

- Data was split into train and test set with 70 / 30 ratio.
- **Accuracy score for the model trained on all features: 0.81**, at the same time precision and recall for churn prediction were rather low (0.69 and 0.57 respectively) which shows potential for model improvement or application of alternative algorithms.
- Confusion matrix for the model is shown in Figure 7.
- Analysis of logit coefficients has shown that features 'tenure', 'Contract_Two year', and 'TotalCharges' have the highest effect in the model.
- By applying recursive feature elimination with step 3 it was possible to reduce number of features from 23 to 7 with minor drop in accuracy to 0.80. The final features used were: **'tenure', 'TotalCharges', 'PhoneService_Yes', 'InternetService_Fiber optic', 'InternetService_No', 'Contract_One year', 'Contract_Two year'.** Evaluation metrics for the resulting model are shown in Figure 8.

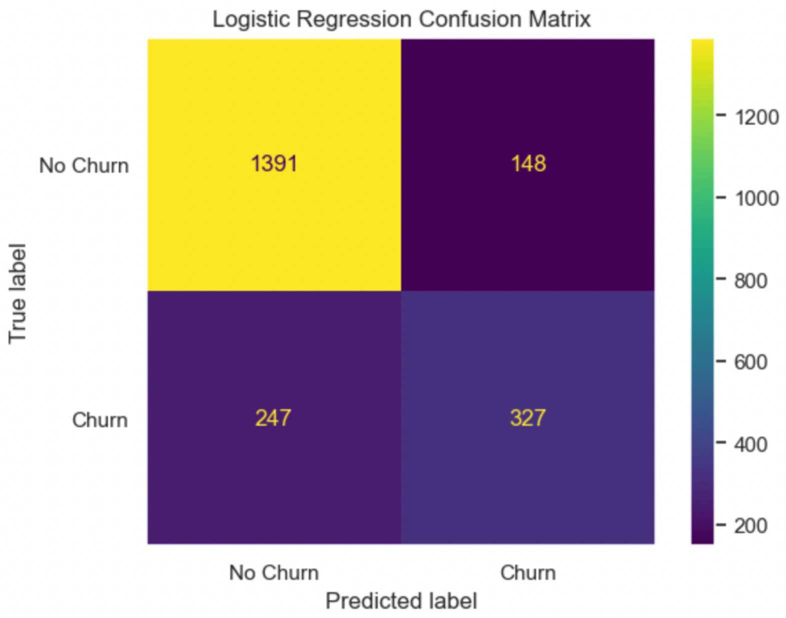
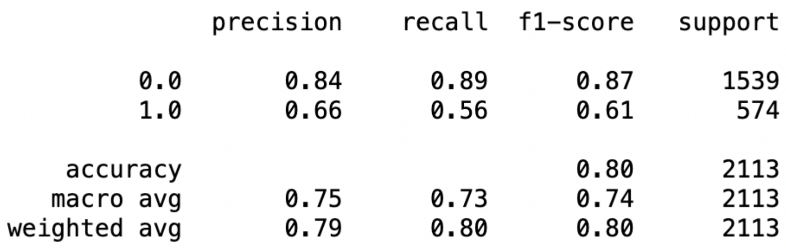


Figure 7 Confusion matrix for logistic regression with all features



```
              precision    recall  f1-score   support

         0.0       0.84      0.89      0.87      1539
         1.0       0.66      0.56      0.61       574

    accuracy                           0.80      2113
   macro avg       0.75      0.73      0.74      2113
weighted avg       0.79      0.80      0.80      2113
```

Figure 8 Final model evaluation metrics

5. Ethical Considerations

As the objective of the ML project was churn prediction among Telecom customers, practical outcome of the model deployment can be targeted interventions (in from of promotions, offers, discounts) which could lead to higher customer satisfaction and as a result churn reduction.

If predictions of the model deem to be inaccurate, the biggest risk is to offer a discount to a wrong customer. This in return can lead to a minor financial loss for the organisation but in longer run result in loyalty and customer retention. For that reason, this project has limited ethical considerations and risks.

There are a few ethical considerations that worth keeping in mind:

a. **The potential ethical concerns surrounding the dataset:**

   i. Selected data set was synthetic and **can't fully reflect real life customer data**. While such data set was suitable for model testing and evaluation it won't be right to bring the model build on such data to production.

   ii. If substituting the data set with real one, it is important to make sure it is **representative of all the firm's customers** and do not contain any potential to develop biases by the algorithm (e.g. for case when location information or postal codes are used customers from lower income areas might be classified differently compared to customers from high income areas). It is important that ML model (and consequently decision and actions that are based on its output) do not put protected and vulnerable groups of customers at risk of different treatment. Despite absence of personally identifiable information (PII), when collecting data clear customer consent and adherence to privacy regulations is critical.

   iii. As the obtained data set does not include any PII, there was **no personal data protection concerns** related to it. Most of demographic features – like age – were aggerated (e.g. under 65 or 65 and older) which is good approach both for the model performance and privacy considerations.

b. **The potential ethical concerns related to pre-processing framework and model**:

   i. As dataset have **minimal amount of missing data** ("TotalCharges", 11 rows out of 7043 or 0.15%) and it was associated with a particular category of customers ("tenure" equals 0) there is no concern for significant altering of dataset with imputing.

   ii. MinMaxScaler was preferred as it **does not change original distribution shape** and reduces risk of in significantly changing data that can lead to incorrect or unexplained outcomes.

   iii. If using PCA or other reduction techniques in the future it is important to consider **model's explainability and transparency of use**. Application of any dimensional reduction technique can make it more compilated to interpret the results and create a perception of "Black box", which is not beneficial and limits control over potential errors or biases.

   iv. While transparency is important for user trust, knowing the factors that lead to churn prediction, some customers might try **to manipulate their usage patterns** to get flagged as potential churners and receive special offers. It is important not to share full details of the model and data that goes into it to prevent gaming the system.

6. Conclusion

To predict churn among Telecom customers an exploratory data analysis was performed on dataset covering 7,043 customers. The dataset was prepared for a ML classifications model including handling missing values, one-hot encoding, and MinMax scaling. Logistic Regression model trained on dataset with 70/30 train-test split and optimised using recursive feature elimination showed overall accuracy of 0.80 but low recall and precision. It shows potential for future feature and model optimisation as well as comparison of performance with other models. Based on the goals of this research, ethical assessment did not find any significant issues that need to be addressed.

Bibliography

[1] V. Lazarov and M. Capota, "Churn Prediction," *Bus. Anal. Course. TUM Comput. Sci.,* no. 33, p. 34, 2007.

[2] Z. Deng, Y. Lu, K. K. Wei and J. Zhang, "Understanding customer satisfaction and loyalty: An empirical study of mobile instant messages in China," *International Journal of Information Management,* vol. 30, no. 4, pp. 289-300, 2010.

[3] H. S. Lee, "Factors influencing customer loyalty of mobile phone service: Empirical evidence from Koreans," *Journal of Internet Banking and Commerce,* vol. 15, no. 2, pp. 1-14, 2010.

[4] B. Lee, The Hidden Wealth of Customers : Realizing the Untapped Value of Your Most Important Asset, Boston, Mass: Harvard Business Review Press, 2012.

[5] M. Alkhurshan and H. Rjoub, "The scope of an integrated analysis of trust switching barriers, customer satisfaction and loyalty," *Journal of Competitiveness,* vol. 12, no. 2, p. 5, 2020.

[6] S.-Y. Hung, D. C. Yen and H.-Y. Wang, "Applying data mining to telecom churn management," *Expert Systems with Applications,* vol. 31, no. 3, pp. 515-524, October 2006.

[7] IBM, "Telco customer churn (11.1.3+)," 11 July 2019. [Online]. Available: https://community.ibm.com/community/user/businessanalytics/blogs/steven-macko/2019/07/11/telco-customer-churn-1113. [Accessed 12 June 2024].

Appendix 1 – Dataset description

Feature name	Non-Null Count	Data Type	Description	Values
customerID	7,043	String	A distinctive ID that identifies each customer.	e.g. 7590-VHVEG
gender	7,043	Categorical (binary)	The customer's gender	Male, Female
SeniorCitizen	7,043	Boolean	Specifies if the customer is 65 or older.	True, False
Partner	7,043	Boolean	Specifies if the customer has a partner.	True, False
Dependents	7,043	Boolean	Specifies if the customer lives with any dependents (children, parents, grandparents, etc.).	True, False
tenure	7,043	Integer	The total number of months that the customer has been with the company.	e.g. 34
PhoneService	7,043	Boolean	Specifies if the customer subscribes to home phone service.	True, False
MultipleLines	7,043	Categorical	Specifies if the customer subscribes to multiple telephone lines.	Yes, No, No phone service
InternetService	7,043	Categorical	Specifies if the customer subscribes to Internet service.	No, DSL, Fiber Optic
OnlineSecurity	7,043	Categorical	Specifies if the customer subscribes to an additional online security service.	Yes, No, No internet service
OnlineBackup	7,043	Categorical	Specifies if the customer subscribes to an additional online backup service.	Yes, No, No internet service
DeviceProtection	7,043	Categorical	Specifies if the customer subscribes to an additional device protection plan for their Internet equipment.	Yes, No, No internet service
TechSupport	7,043	Categorical	Specifies if the customer subscribes to an additional technical support plan with reduced wait times.	Yes, No, No internet service.
StreamingTV	7,043	Categorical	Specifies if the customer uses their Internet service to stream television programming from a third-party provider (no additional fee).	Yes, No, No internet service
StreamingMovies	7,043	Categorical	Specifies if the customer uses their Internet service to stream movies from a third-party provider (no additional fee).	Yes, No, No internet service
Contract	7,043	Categorical	Customer's current contract type.	Month-to-Month, One Year, Two Year
PaperlessBilling	7,043	Boolean	Specifies if the customer has chosen paperless billing.	True, False
PaymentMethod	7,043	Categorical	Specifies how the customer pays their bill.	Electronic Check, Mailed Check, Credit Card (automatic), Bank transfer (automatic)
MonthlyCharges	7,043	Float	Customer's current total monthly charge for all their services from the company.	e.g. 56.95
TotalCharges	7,043	Float	Customer's total charges, calculated to for their tenure (tenure x MonthlyCharges).	e.g. 1889.5
Churn	7,043	Categorical (binary)	Yes = the customer left the company. No = the customer remained with the company.	Yes, No