Supervised Machine Learning

TELCO CUSTOMER CHURN Data

Outline

1. Objective
2. Dataset Information
3. Data Preprocessing
4. Exploratory Data Analysis
5. Develop Machine Learning Model
6. Conclusion

Outline

- The analysis aims to find the best model for investigating factors influences churn decision of telecommunication customers.

Dataset Information

The dataset contains information as in the following

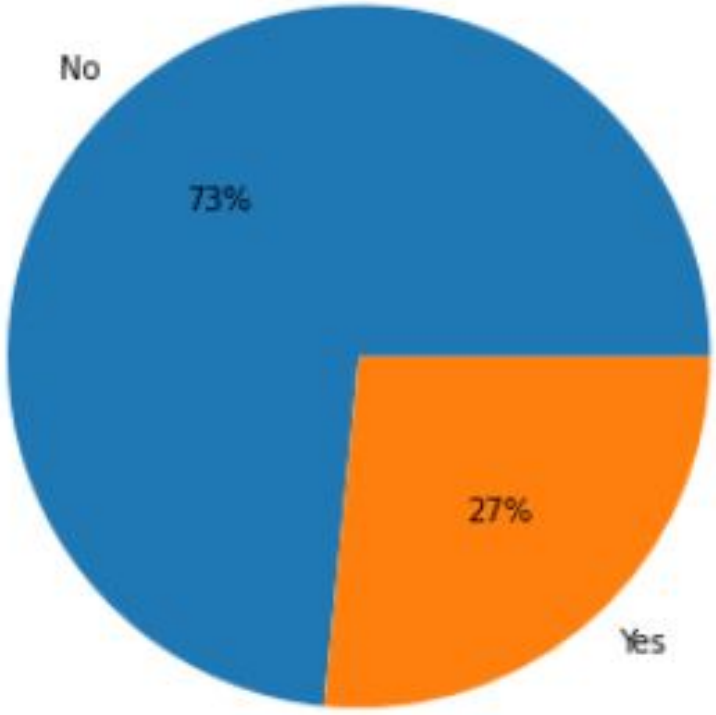
Features				Target
Customers' demography	**Telco Services**	**Charge**	**Related to contract**	**Churn**
Gender Senior citizen Partner Dependents Tenure	Phones service Multiple lines Internet service Online security Online backup Device protection Tech support Streaming tv Streaming movies	Total charges Monthly charges	Payment method Paperless billing Contract	Churn

Data Preprocessing

DATA CLEANING

- In the process of cleaning data, my findings are:
  - Missing values are found only in totalcharges. The variable's type is still an object, but due to missing value, it cannot be convert to integer. After removing the missing values which were 11, the variable can be conver
  - No duplicates

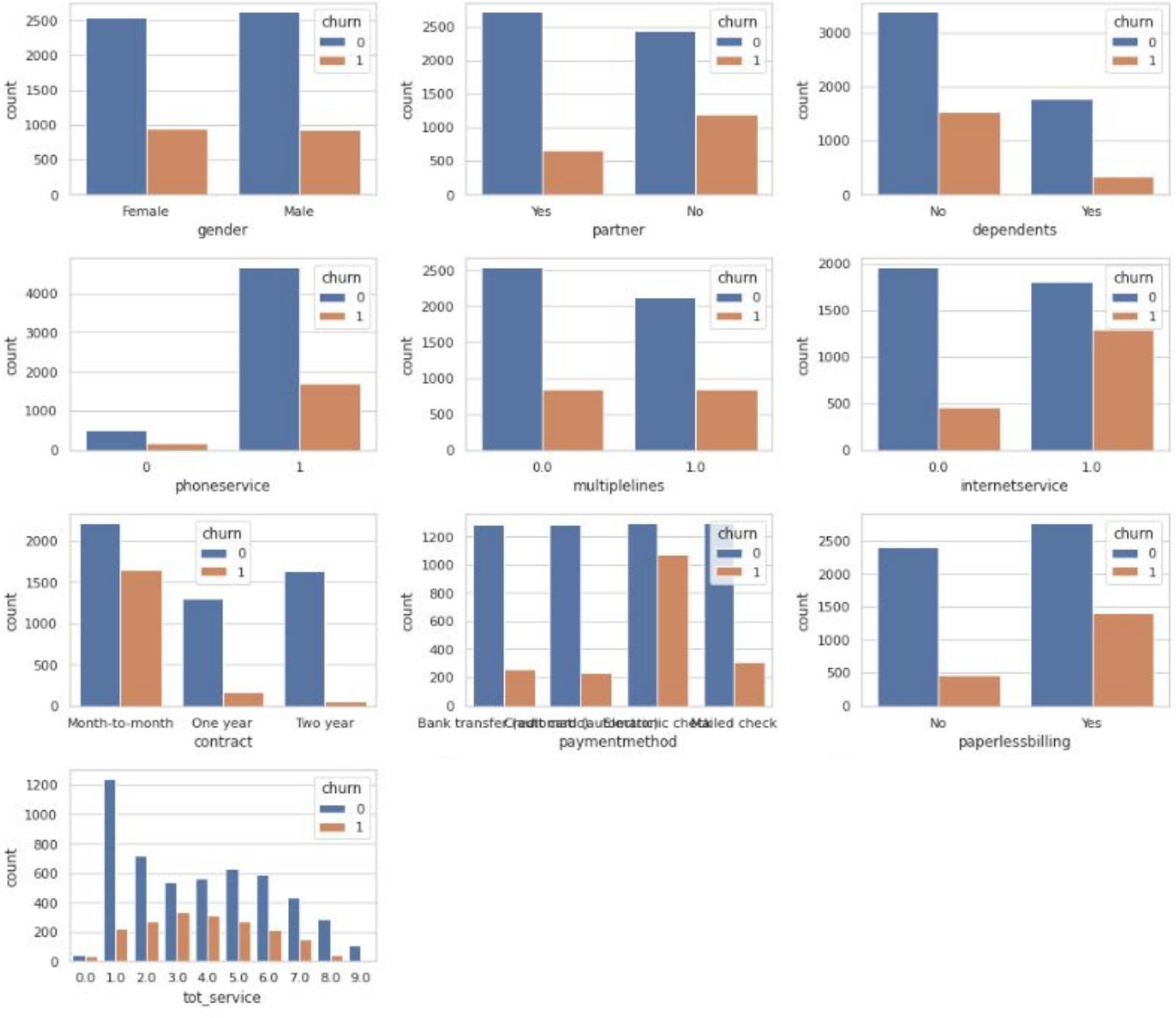
Findings



Churn Rate is pretty high, reaching 27% or equivalent 1899

Churn yes

Churn no



Findings: Customers' Characteristics

Customers who stop using the service have characteristics as follows:

- No partner
- No dependents
- Contract on a month-to basis with paperless billing
- Use bank transfers and check as the payment method
- Use many services (more than 3)

Findings: Summary Statistics

	tenure	monthlycharges	totalcharges	tot_service
count	7032.000000	7032.000000	7032.000000	7032.000000
mean	32.421786	64.798208	2283.300441	3.803612
std	24.545260	30.085974	2266.771362	2.284171
min	1.000000	18.250000	18.800000	0.000000
25%	9.000000	35.587500	401.450000	2.000000
50%	29.000000	70.350000	1397.475000	4.000000
75%	55.000000	89.862500	3794.737500	6.000000
max	72.000000	118.750000	8684.800000	9.000000

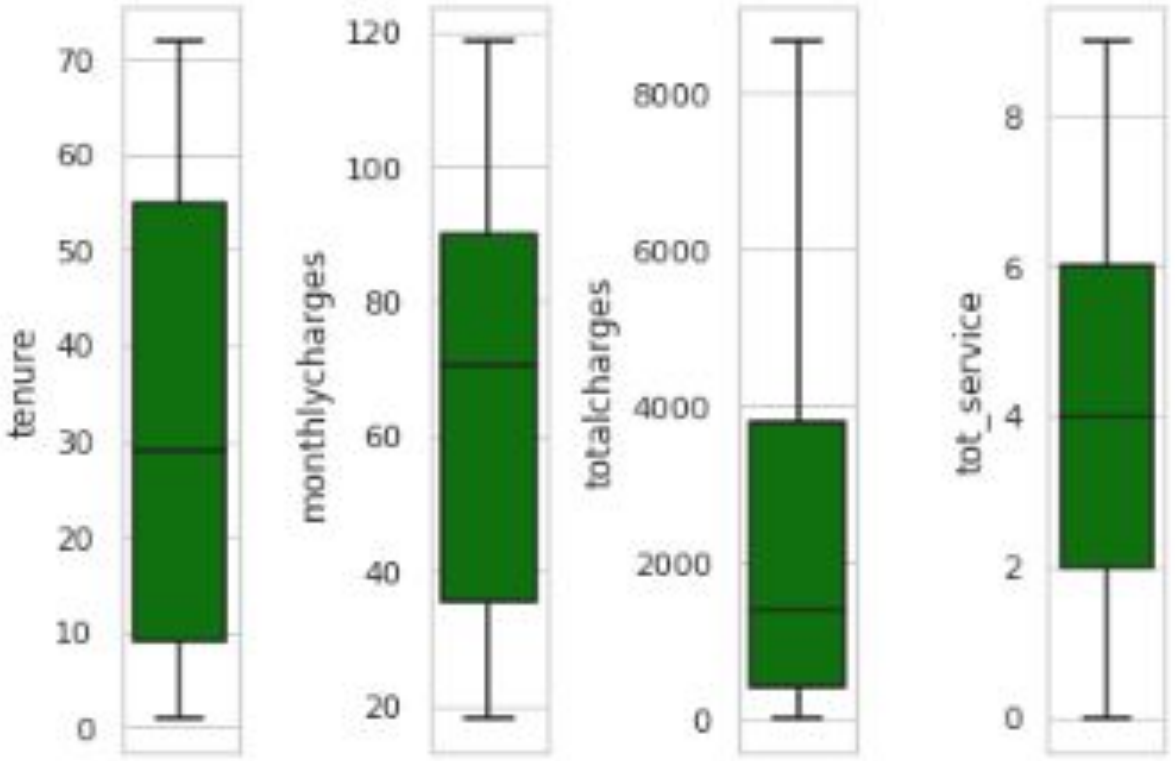
The summary statistics show that the variables are not normally distributed. The Kernel distribution also shows similar result. However, we consider this as normal, and thus proceed the process

Findings: Summary Statistics

	phoneservice	multiplelines	internetservice	onlinesecurity	onlinebackup	deviceprotection	techsupport	streamingtv	streamingmovies	contract	paperlessbilling	paymentmethod
count	7032	6352.0	5512.0	5512.0	5512.0	5512.0	5512.0	5512.0	5512.0	7032	7032	7032
unique	2	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	3	2	4
top	1	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	Month-to-month	Yes	Electronic check
freq	6352	3385.0	3096.0	3497.0	3087.0	3094.0	3472.0	2809.0	2781.0	3875	4168	2365

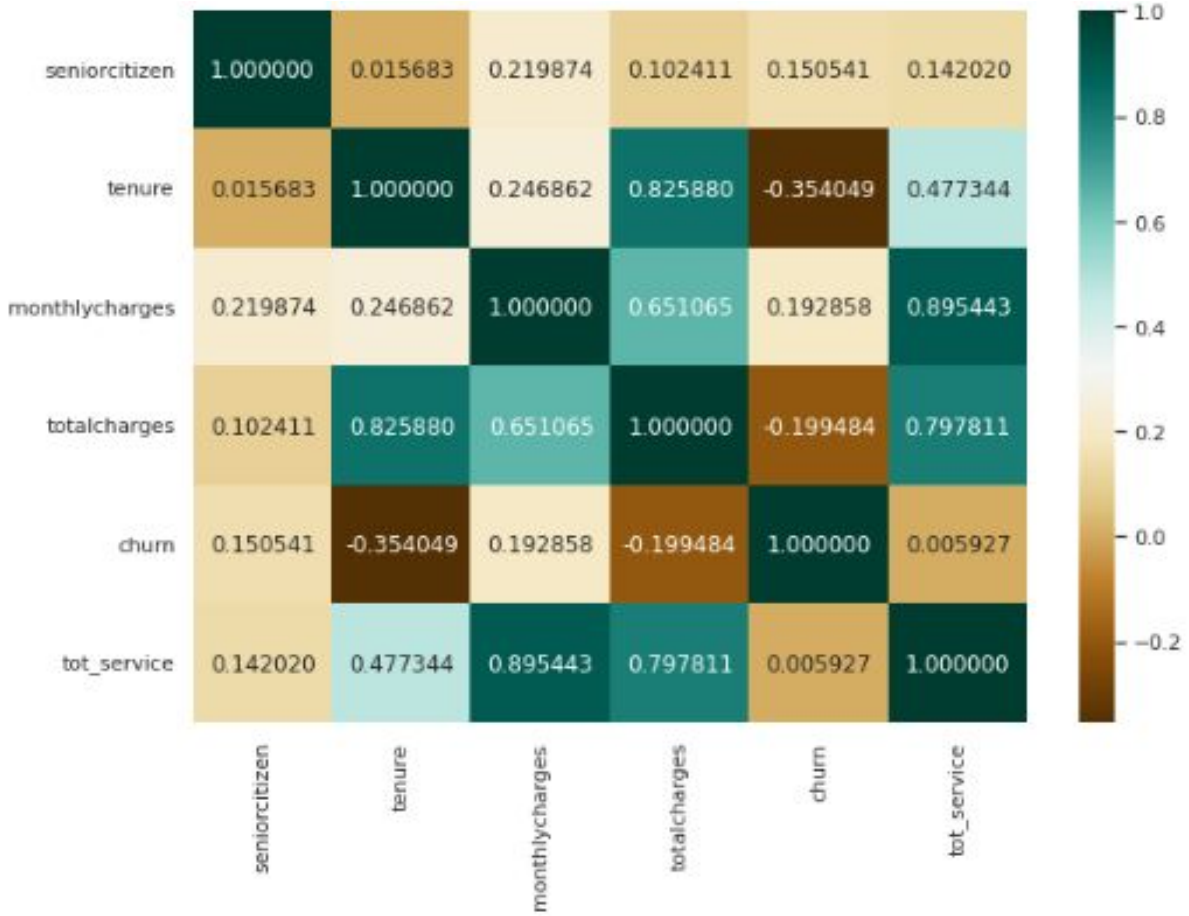
Services that are used the most are phoneservice and internetservice

Findings: Univariate Analysis

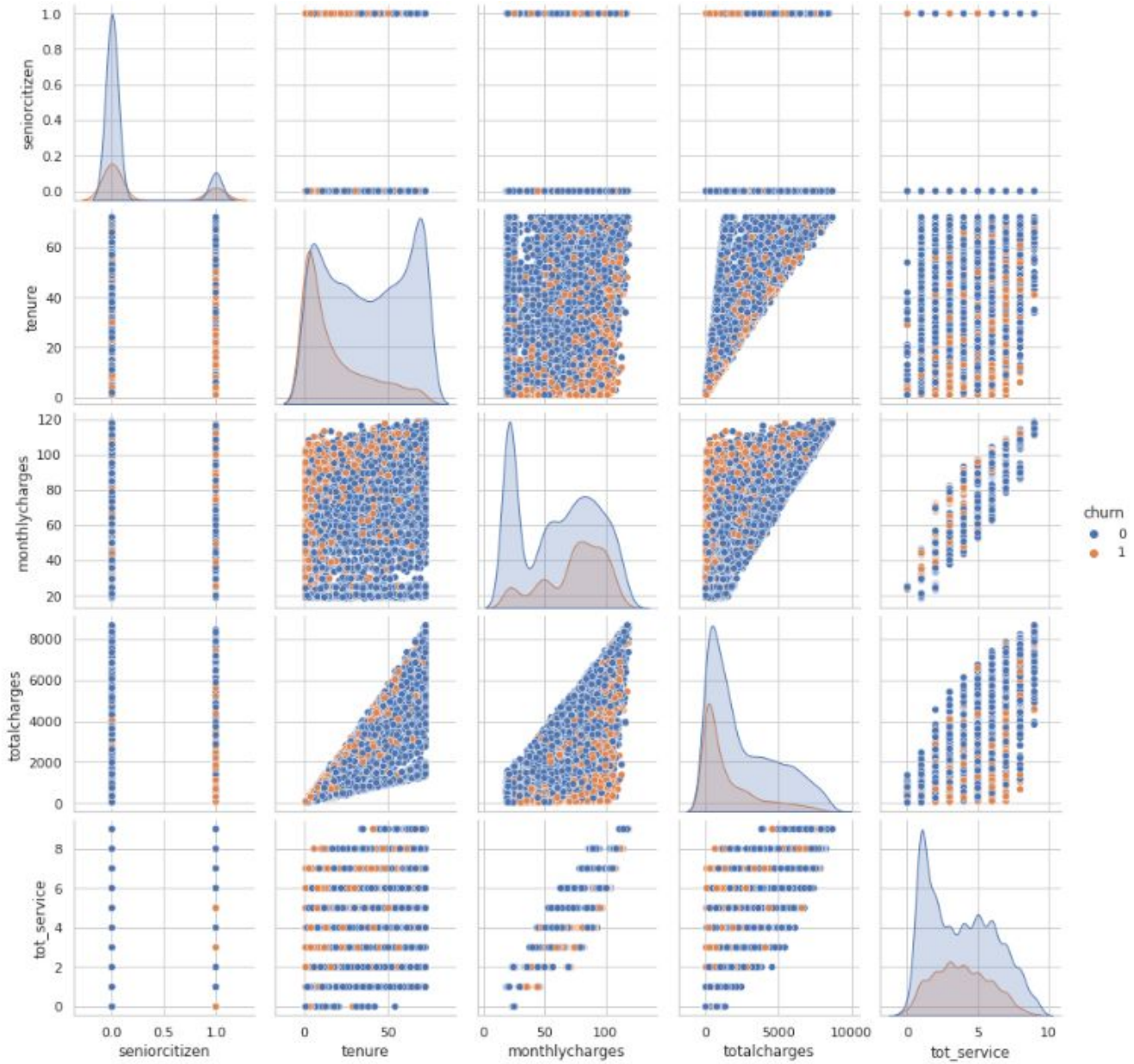


No outliers

Findings: Correlation



- tot_service and monthlycharges, as well as tot_service and totalcharges have a strong correlation, which is reasonable
- totalcharges and tenure also have pretty high correlation
- interestingly, churn has a negative and mild correlation with tenure



Findings

- I used churn as the hue parameter to explain the reasons people stop using the service
- It seems customers who stop using the service are likely to be a senior citizens with relatively short years of working and relatively high bills

Deep Dive: Characteristics of Customers who Stop Using the Service

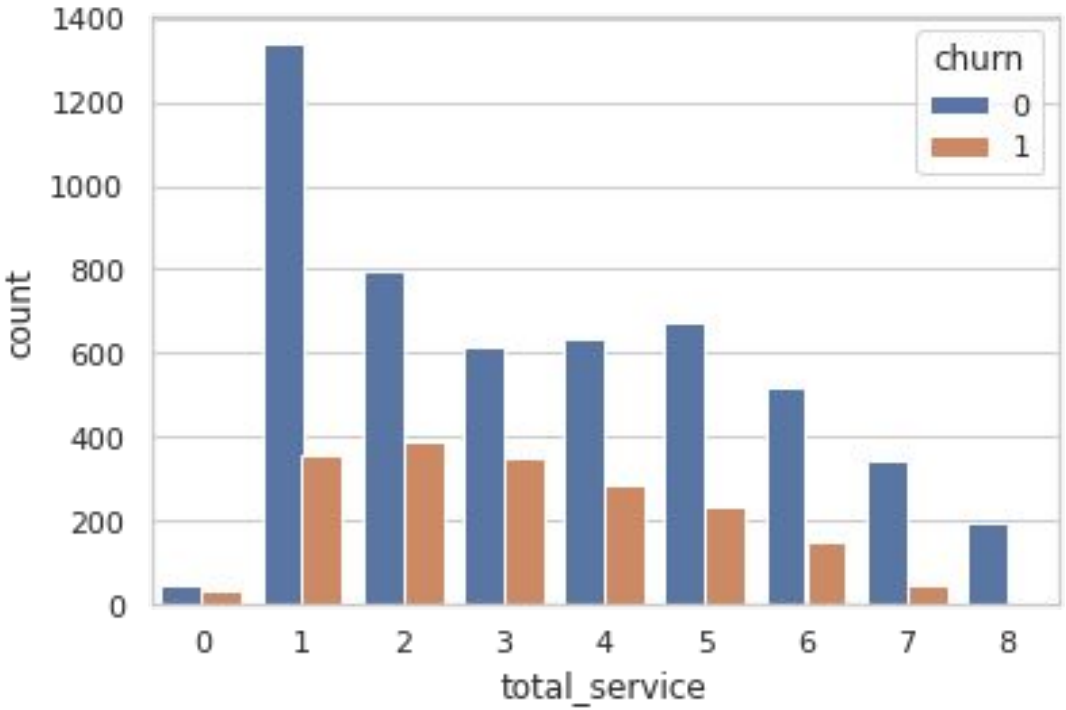
	monthlycharges								seniorcitizen		...	tot_service		totalcharges							
gender	Female				Male				Female		...	Male		Female				Male			
partner	No		Yes		No		Yes		No		...	Yes		No		Yes		No		Yes	
dependents	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	...	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
churn																					
0	59.669663	52.504911	71.630097	57.675469	57.074197	46.805294	73.170650	59.756309	0.148876	0.017857	...	4.730081	3.928859	1978.217135	1359.351339	3623.278964	2808.451475	1776.013682	1540.191176	3648.989593	2891.177651
1	73.475213	67.748485	81.432620	73.143939	70.103918	63.631818	83.446137	77.468803	0.269165	0.030303	...	4.609442	4.316239	1079.102385	944.440909	2311.314973	1941.395455	1124.731530	893.650000	2546.501502	2344.748718

2 rows × 40 columns

```
churn  gender  partner  dependents
0      Female  No       No            1068
                        Yes            112
               Yes      No             618
                        Yes            746
       Male    No       No            1089
                        Yes            170
               Yes      No             615
                        Yes            745
1      Female  No       No             587
                        Yes             33
               Yes      No             187
                        Yes            132
       Male    No       No             536
                        Yes             44
               Yes      No             233
                        Yes            117
dtype: int64
```

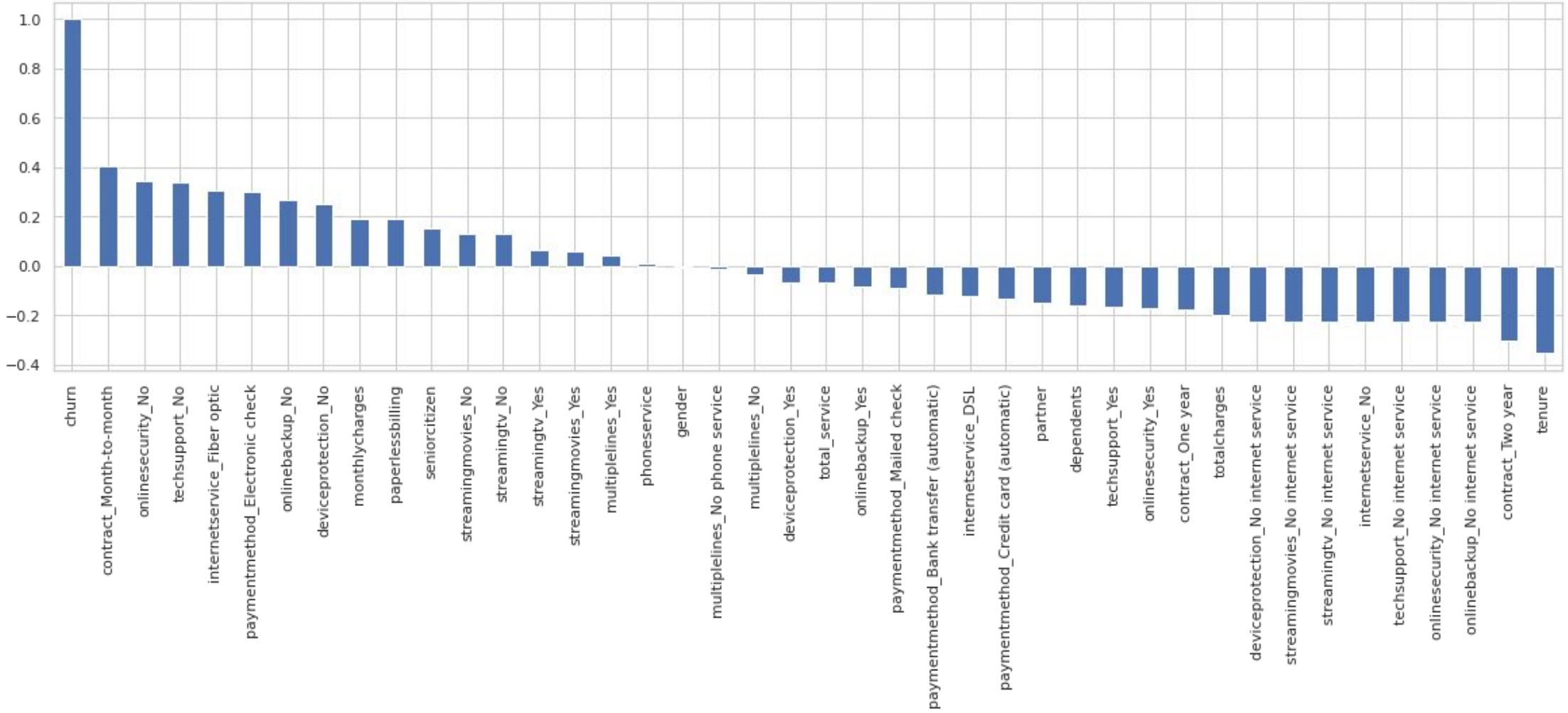
Across characteristics, those who stop using the service have relatively higher monthly bills and shorter years of tenure, ni partner, and have relatively higher monthly charges than those who keep using the service.

Deep Dive: Does having one more service trigger churn decision?



It seems that having more service has relation to churn decision

Contract type and tenure have association with churn decision



Supervised Machine Learning

KNN

```
Confusion Matrix  [[897 131]
 [200 179]]
AUC  0.6724318039485437
F1 Score  0.5195936139332367
Precision  0.5774193548387097
Recall  0.47229551451187335
```

Logistic Regression

```
Confusion Matrix  [[924 104]
 [175 204]]
AUC  0.7185456300113959
F1 Score  0.593886462882096
Precision  0.6623376623376623
Recall  0.5382585751978892
```

Decision Tree

```
Confusion Matrix  [[812 216]
 [183 196]]
AUC  0.6535168321304271
F1 Score  0.49557522123893805
Precision  0.47572815533980584
Recall  0.5171503957783641
```

Random Forest

```
Confusion Matrix  [[915 113]
 [195 184]]
AUC  0.6877829738303748
F1 Score  0.5443786982248521
Precision  0.6195286195286195
Recall  0.48548812664907653
```

The metrics of Logistic Regression is the highest among all, and thus we will use this model to analyze the data further

Conclusion

Conclusion

Our findings suggest that the characteristics of customers who stop using the service are:

- No partner
- No dependents
- Contract on a month-to basis with paperless billing
- Use bank transfers and check as the payment method
- Use many services (more than 3)

In addition, customers who stop being loyal have relatively higher monthly bills and shorter years of tenure and have relatively higher monthly charges than those who keep using the service.

Further findings:

- It seems that having more service has relation to churn decision
- Contract type and tenure have association with churn decision

Recommendation

- **Recommendation**
- Develop a contract method or service package that is in line with the needs of young customers
    - In addition, careful to offer services other than telephone and internet service to young customers
- Improve the payment method, aiming to provide a payment method that is less hustle for customers
- Based on 4 models, Logistic Regression has the highest evaluation metrics among all other models. Therefore, we can estimate churn decision by using this model