Data Mining mini projest

Katayoon Sadeghi mehr Feb 2020

Logistic Regression/Decision Tree: Churn Costumers

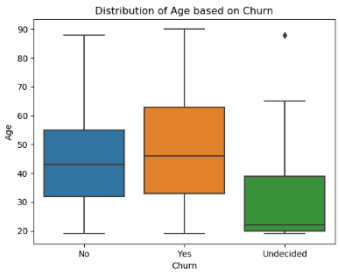
The dataset of churn contains 14 columns. **customerID**, **tenure**, **PhoneService**, **Contract** , **Paperless**, **PaymentMethod**, **MonthlyCharge**, **MontlyGroceryspent**, **Revenue**, **gender**, **Age**, **Partner**, **N_dependent** and **Churn** as dependent variable.
Monthly Charges has 12, churn 52 and revenue 44 missing values.
median is used to replace revenue and monthly charges missing values, and "Undecided" for the 52 missing values for y=churn.
After cleaning we have **7042** rows of data.
Encoding on categorical variables, Phone Service with 2 levels yes=1, no=0, Contract with 3 levels month-to-month=1, one-year=2 and two-year=3, Paperless with 2 levels yes=1, no=0, PaymentMethod with 4 levels Bank transfer (automatic)=1, Credit card (automatic)=2, Electronic check=3 and Mailed check=4, gender with 2 levels male=1 and female=2, Partner with 2 levels yes=1, no=0, churn with 3 levels undecided=1, no=2 and yes=3 is done.

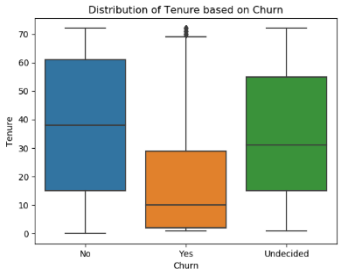
customerID	tenure	Phone Service	Contract	Paperless	Payment Method	Monthly Charges	Monthly Grocery spent	Churn	Revenue	gender	Age	Partner	Dependents	N_ dependent
6823-SIDFQ	28	Yes	One year	No	Credit card (automatic)	18.25	259	No	22024	Male	53	No	No	0

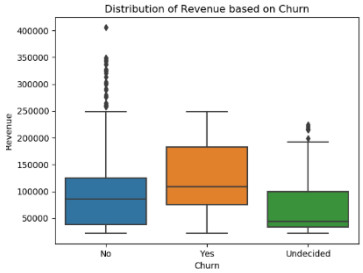
Distribution of variables:

Average of age, tenure and revenue for different levels of churn. Apparently, the undecided ones are from younger ages. Those who decided to churn have lower tenure and those who do not want to churn are mostly from lower income with some outliers of high incomes.

Churn	Age	tenure	Revenue
Yes	47.65	17.95	117474.57
No	44.48	37.61	95520.22
Undecided	29.19	34.49	73489.66

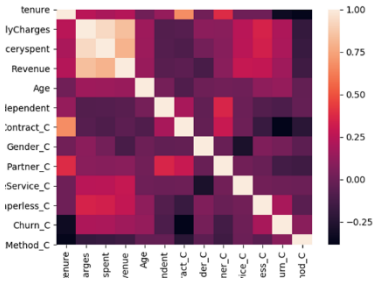






Correlation and association between variables:

There is pretty high correlation between y=churn and some dependent variables. Also some of the dependent variables are highly correlated with each other.



	tenure	Monthly Charges	Monthly Grocery spent	Revenue	Age	N_ dependent	Contract_C	Gender_C	Partner_C	Phone Service_C	Paperless_C	Churn_C	Payment Method_C
tenure	1												
MonthlyCharges	0.2470	1											
MonthlyGroceryspent	0.1999	**0.9310**	1										
Revenue	0.2309	**0.8384**	**0.7980**	1									
Age	0.0149	0.1593	0.1522	0.1452	1								
N_dependent	0.1343	-0.0831	-0.0778	-0.0658	0.0295	1							
Contract_C	**0.6715**	-0.0750	-0.1018	-0.0526	-0.0927	0.1922	1						
Gender_C	0.0173	0.0999	0.0261	-0.1227	0.0054	-0.0329	-0.0304	1					
Partner_C	0.3800	0.0957	0.0841	0.0943	0.0187	0.3581	0.2951	-0.0154	1				
PhoneService_C	0.0084	0.2474	0.2504	0.2867	0.0065	-0.0041	0.0022	-0.2824	0.0178	1			
Paperless_C	0.0060	0.3510	0.3351	0.2777	0.1130	-0.0824	-0.1770	0.0645	-0.0148	0.0165	1		
Churn_C	-0.3321	0.2058	0.2033	0.1624	0.1320	-0.1074	-0.3835	0.0275	-0.1353	0.0056	0.1933	1	
PaymentMethod_C	-0.3703	-0.1919	-0.1745	-0.1530	-0.0302	-0.0258	-0.2273	-0.0491	-0.1550	-0.0041	-0.0627	0.0945	1

Logistic Regression:

X = tenure, MonthlyCharges, MontlyGroceryspent, Revenue, Age, N_dependent, Paperless_C, PaymentMethod_C, Contract_C, Gender_C, Partner_C, PhoneService_C and

y = churn

By removing the variables that are not significant in the model. The final model is:

	Churn=1	Churn=2		Churn=3		
X	Coef	Coef	P>|z|	Coef	P>|z|	Min p
Intercept	-5.052	1.335	0.000	-1.644	0.000	0.000
tenure	0.145	0.813	0.000	-0.929	0.000	0.000
Monthly Charges	-0.617	-0.712	0.000	0.942	0.000	0.000
Age	-1.337	-0.005	0.583	0.092	0.006	0.006
N_dependent	-0.472	0.145	0.010	-0.098	0.433	0.010
Contract_C	0.160	0.572	0.002	-0.768	0.000	0.000
Gender_C	-0.343	0.100	0.001	-0.031	0.791	0.001
PhoneService_C	0.107	0.194	0.091	-0.273	0.000	0.000

y	1	2	3
1	0	29	5
2	0	1150	102
3	0	250	225

	precesion	recall	f1-score	support
1	0	0	0	34
2	0.80	0.92	0.86	1252
3	0.68	0.47	0.56	475
Accuracy			**0.78**	1761
macro Avg	0.49	0.46	0.47	1761
weighted Avg	0.75	0.78	0.76	1761

Accuracy for cross validation, having **7-fold** is **0.7761**

Decision Tree:

X = tenure, MonthlyCharges, MontlyGroceryspent, Revenue, Age, N_dependent, Paperless_C, PaymentMethod_C, Contract_C, Gender_C, Partner_C, PhoneService_C and

y = churn with max depth of 5

y	1	2	3
1	3	30	10
2	3	1311	218
3	0	229	309

	precesion	recall	f1-score	support
1	0.5	0.07	0.12	43
2	0.84	0.86	0.85	1532
3	0.58	0.57	0.57	538
Accuracy			**0.77**	2113