PREDICTING CUSTOMER CHURN WITH MACHINE LEARNING

by Leo Evancie

Agenda

Problem and solution

Data science method

Data wrangling
Exploratory analysis
Preprocessing
Modeling

Conclusions and next steps

The Problem

- Customer acquisition costs more than customer retention
- Some customers leave, or “churn”
- Failing to predict churn -> loss of revenue

The Solution

- Apply the data science method
- Analyze thousands of telecom customers (demo and behavior)
- Logistic regression model to classify 1 (“churn”) or 0 (“no churn”)
- Churn analysis and prediction could:
  - *Guide interventions*
  - *Shape sales and marketing*

DATA WRANGLING

Raw Data

- Over 7,000 TelCo customers (IBM sample dataset via Kaggle)
- Demographic, behavioral, and purchasing
- Minimal data quality issues

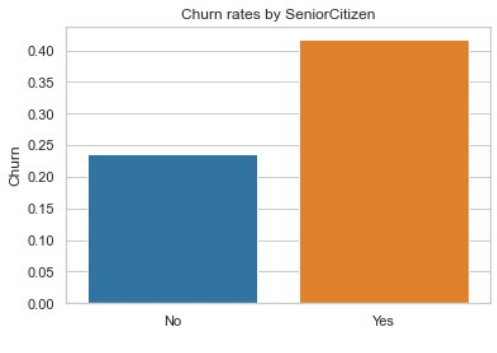
#	Column	Non-Null Count	Dtype
0	customerID	7043 non-null	object
1	gender	7043 non-null	object
2	SeniorCitizen	7043 non-null	int64
3	Partner	7043 non-null	object
4	Dependents	7043 non-null	object
5	tenure	7043 non-null	int64
6	PhoneService	7043 non-null	object
7	MultipleLines	7043 non-null	object
8	InternetService	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

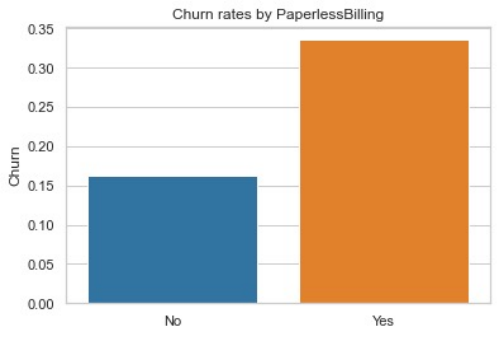
EXPLORATORY ANALYSIS

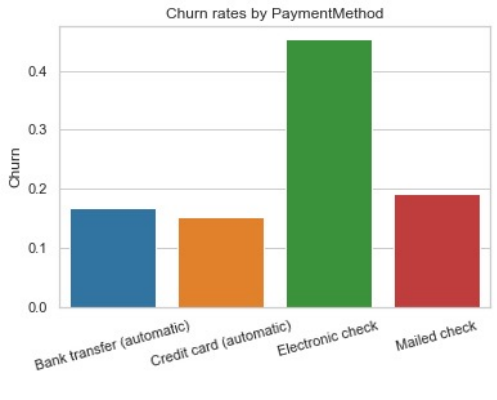
Target variable

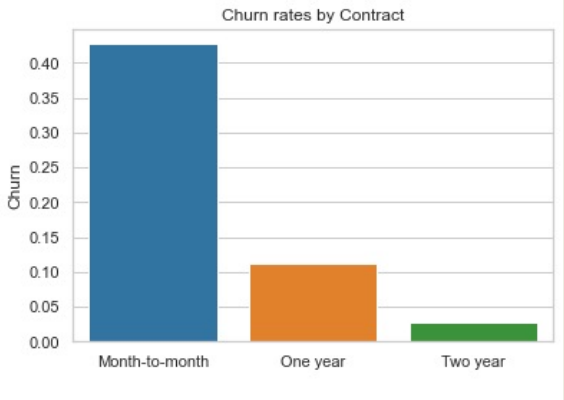
- 26.65% of 7,000 customers churned

Categorical features



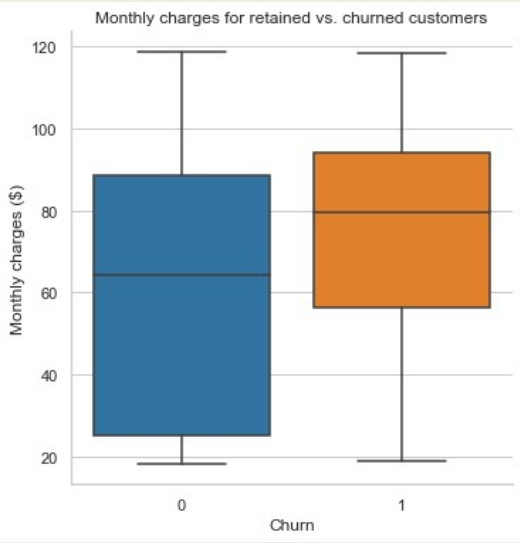


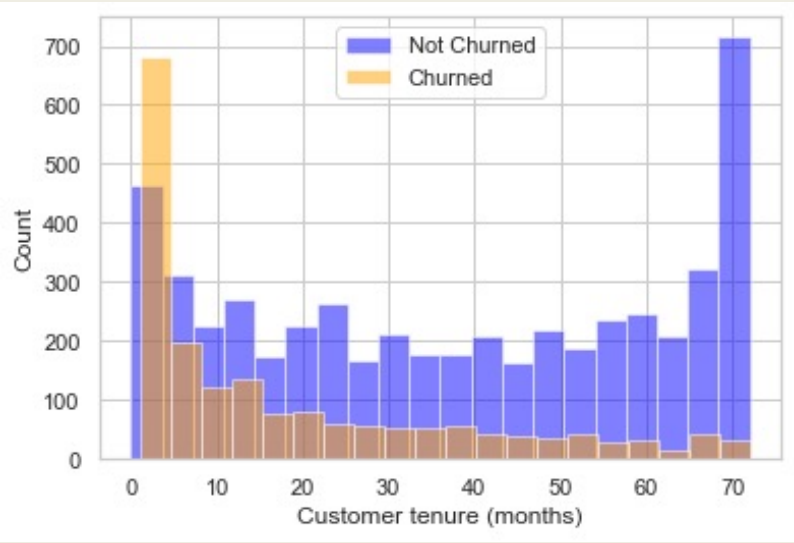




- Showing four of several key features
- PhoneService and MultipleLines showed no clear relationship to churn

Numerical features





PREPROCESSING

1

Encode ‘Churn’ as 0 or 1

2

Create dummies for categorical features

3

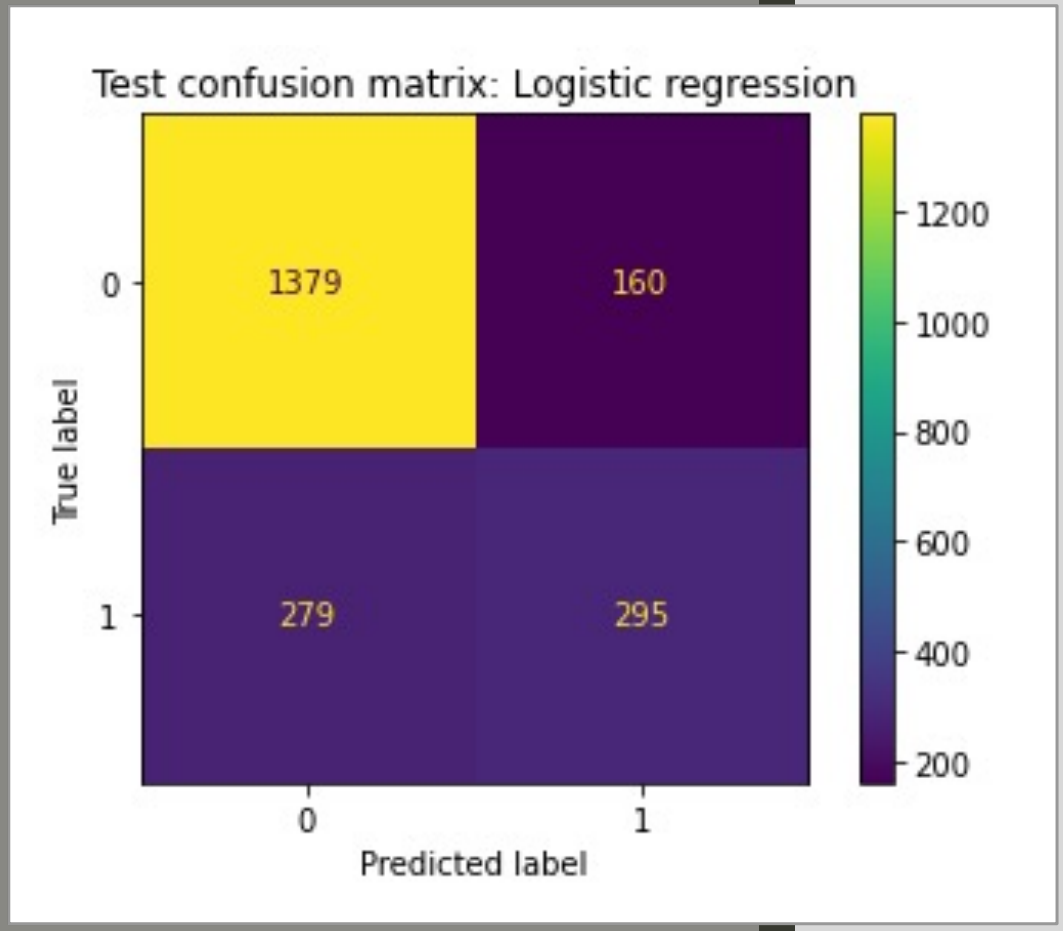
Train/test split

4

Rescale numerical features to between 0 and 1

MODELING

- Built models with default parameters for initial comparison:
  - *Logistic regression*
  - *Random forest classifier*
  - *Support vector classifier*
- Hyperparameter tuning for LR model
- Default LR outperformed tuned LR:
  - *penalty=l2, C=1.0, solver='lbfgs'*



training scores

	precision	recall	f1-score	support
0	0.85	0.90	0.87	3635
1	0.66	0.56	0.61	1295
accuracy			0.81	4930
macro avg	0.76	0.73	0.74	4930
weighted avg	0.80	0.81	0.80	4930

testing scores

	precision	recall	f1-score	support
0	0.83	0.90	0.86	1539
1	0.65	0.51	0.57	574
accuracy			0.79	2113
macro avg	0.74	0.70	0.72	2113
weighted avg	0.78	0.79	0.78	2113

CONCLUSIONS AND NEXT STEPS

- Bivariate analyses revealed patterns between churn and certain variables
  - *Churn happens early in tenure*
  - *Senior citizens are more likely to churn, however,*
  - *Customers using more internet-based services also churned more often*
  - *Lower monthly costs associated with less churn*
  - *Customers paying month-to-month were far likelier to churn*
- Qualified success using logistic regression model to predict churn
- Even limited predictive power could save resources: Targeted intervention to at-risk customers and strategic promotion of low-churn services and products