Telecom Churn Prediction

Kyle Lawrence

May 24, 2018

Class: Mar 20 Part Time Data Science

Agenda

- Problem statement
- Data Source & Features
- Exploratory Data Analysis
- Model Selection
- Conclusions

Problem Statement

How can we predict customer churn in the telecom industry?

- A customer **churns** when they leave a product/service
- Customer churn is important to understand:

    Customer retention costs are high

    Customer acquisition is even higher [1]

    Churn bleeds revenue to competitors

- Annual rate of churn in telecom at 30% and growing as competition increases [1]

Data Source & Features

Retreived from Kaggle
Before cleansing: 7043 rows, 21 columns

Features include:

Product attributes

Customer attributes

Billing attributes

TARGET: churned (Y/N)

Dummy conversion:

has_internet_service
3 classes

contract
3 classes

payment_method
4 classes

Feature engineering:

Number of products

Hierarchical clustering
2 clusters
Silhouette score = 0.371

Post engineering: 7032 records, 25 features

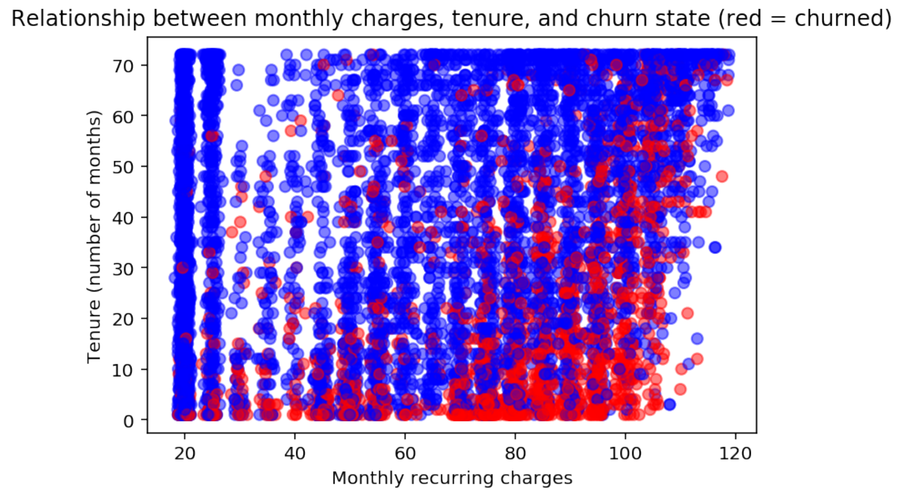
Exploratory Data Analysis

- 26% of dataset churned
- Sex is insignificant — remove this feature
- 90% have phone service
- Most churners had 2 to 5 products
- Lot's of multicollinearity between features, particularly problematic:

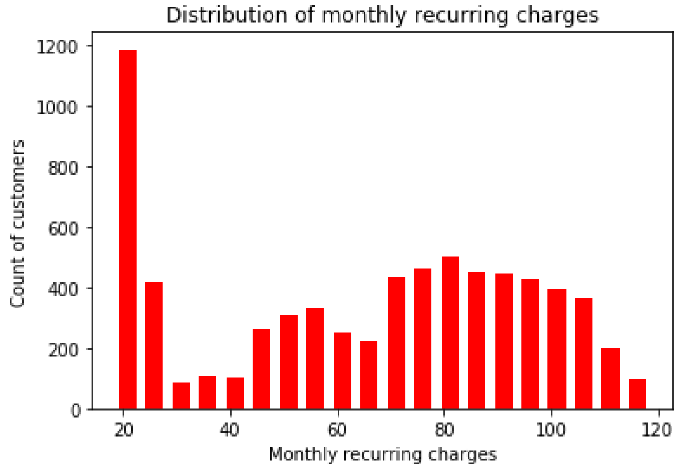
  total_charges

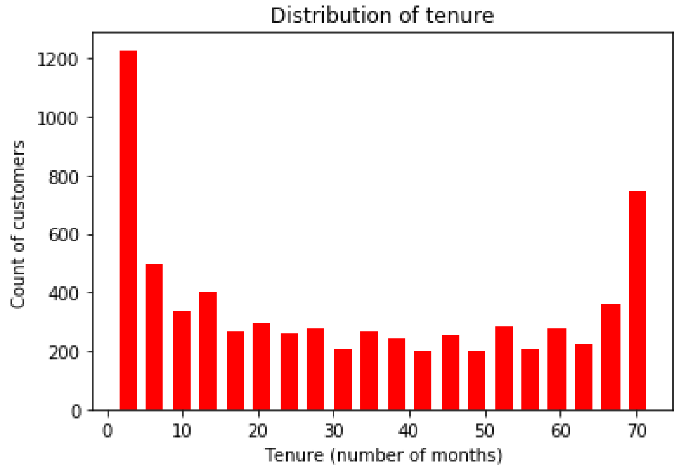
  num_products (doh!)

  remove these features
- Generally low correlation between target and features

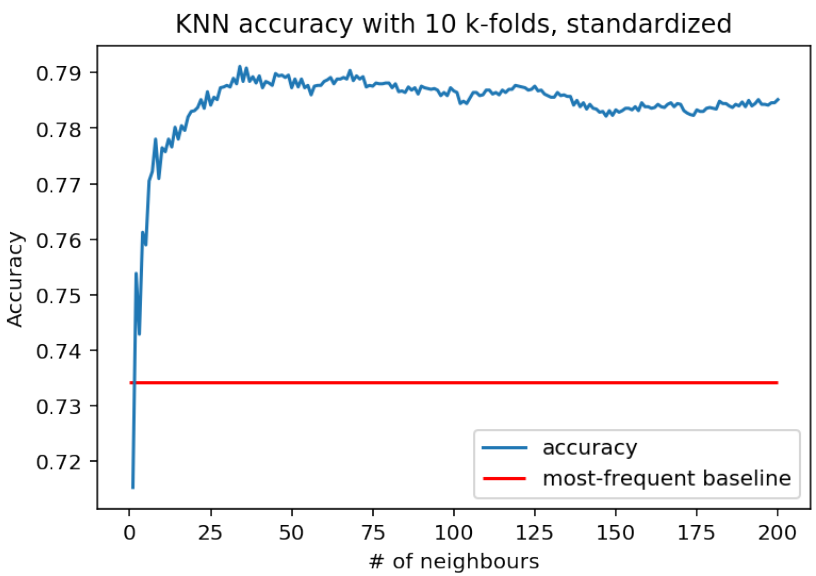


Some non-normal distributions





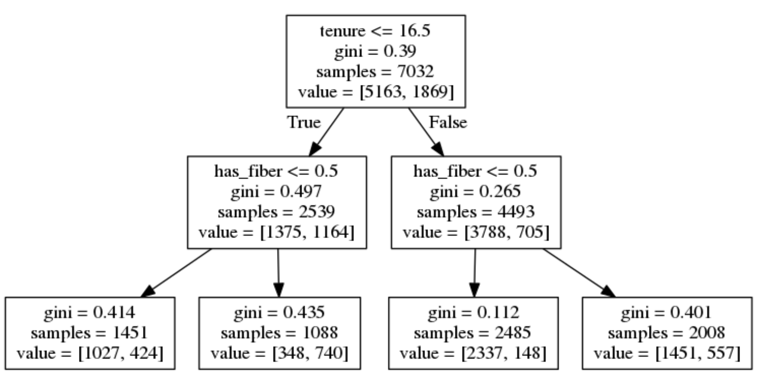
Modelling



K-Nearest-Neighbours

Why: good with irregular decision boundaries, robust on standardized data

- Standardized through sklearn.pipeline
- 22 features
- Score=0.791, 5.7% better than baseline
- Best k=34



feature	importance
tenure	0.502096
has_fiber	0.497904
is_senior_citizen	0.000000
has_paperless_billing	0.000000
payment_method_mail	0.000000

Decision Tree (DT)

Why: not impacted by feature irrelevance, interpretable, fast

- 22 features
- Score=0.789, 5.5% better than baseline
- Best depth = 2
- Grid search across 3 params: only depth matters
- Gini coefficient only good on one node
- Tenure and has_fiber are the only important features

Modelling (cont’d)

```
rf_grid.best_score_
```

```
0.7943686006825939
```

```
rf_grid.best_params_
```

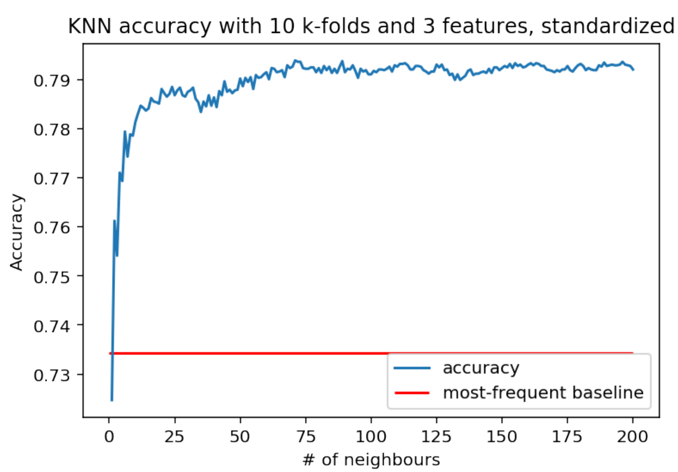
```
{'max_features': 7, 'n_estimators': 180}
```

feature	importance
tenure	0.261436
monthly_charges	0.257965
contract_two_yr	0.042938
payment_method_echeck	0.042886
has_fiber	0.041077

Random Forest Classifier (RFC)

Why: better accuracy, competitive with best supervised learning

- 22 features
- Best score = 0.794
- Grid search yields max features=7, n_estimators=180
- Tenure and monthly_charges important, but worse than decision tree
- 0.5% better than decision tree



K Nearest Neighbours, 3 features

Why: KNN does well absent of feature irrelevance

- Based on DT, RFC: tenure, monthly_charges, has_fiber
- Score = 0.794
- Best k = 71
- Negligibly better than high-dimension KNN

Conclusions

- None of the models performed much better than the baseline – only about 6% more accurate than predicting no churn for everyone
- Next steps:

1) Normalize distributions
2) More dimensionality reduction
3) Possible rebalancing of some classes
4) Investigate other models (logistic & linear classifier)

References

- [1] https://ieeexplore.ieee.org/document/6340176/