Telecom Customer Churn – Report

The Problem

Businesses need customers, and gaining them can be difficult and expensive. That's why retaining customers once won is so important, and reducing churn, when customers choose to stop being customers, is maybe more important.

The Telco dataset is a fictional yet real-world representative dataset containing information on customers and whether they churned. I use it here to practice excellent data science.

The Data

One CSV file:

- Customer ID (All unique)
- Gender (M/F)
- Senior Citizen (T/F)
- Partner (Yes/No)
- Dependents (Yes/No)
- Tenure (In years)
- Phone Service (Yes/No)
- Monthly Charges (In $)
- Payment Method (Electronic Check, Bank Transfer…)
- Total Charges (In $)
- Churn (Yes/No)
- Etc.

The Goal

Understand why customers are churning and also predict who will churn.

The Metrics of Success

Accuracy and Recall Score

To measure the success of this model *of course* accuracy is relevant, but false negatives (customers predicted as *not going to churn*, but they do churn) are particularly painful for the business in this case. With this, recall is a relevant metric to account for the ratio of false negatives.

The Approach

Firstly, import, clean, explore, visualize, and understand the data before jumping in and throwing ML at the dataset. Visually explore single features, then relationships between features, then correlation with the target variable (churn), etc.

Next, segment customers into intuitive groups so to understand what's happening within them. For example, low, mid, and high spending groups regarding monthly charges in dollars. Take the average of each and voila, insight.

Transform the data into usable formats for ML algorithms and engineer any features which may help predict churn.

Finally, model the problem with ML algorithms. This is a classification problem to binarily classify customers as churned / not churned; 0 is good 1 is bad. Experiment with many untuned classification ML algorithms and see which attain the best accuracy and recall scores. While these metrics are important goals, being able to derive insights from the models is also necessary for the business, so algorithms which can provide the importance of features for their predictive power are more valuable.

Then, refine some of the best models with grid-search cross-validation (hereafter GSCV) to find the optimal hyperparameters. Of the refined models, choose the best performer on a basis the goal metrics.

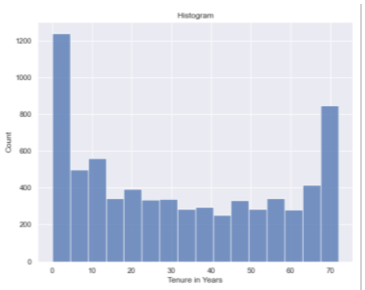
Finally, commit to the best model, analyze the results, and derive business insight.

The Findings

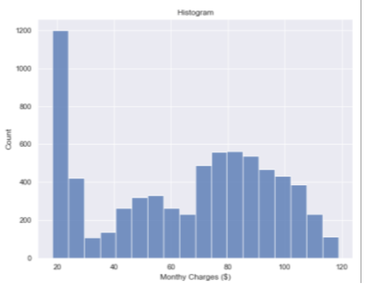
The data generally is very clean, especially with no missing values. The only cleaning necessary was to swap spaces in Total Charges with 0s so to convert the column to numeric. The rest was converting features from string type to numeric type, easy tasks.

Interesting characteristics of the data:

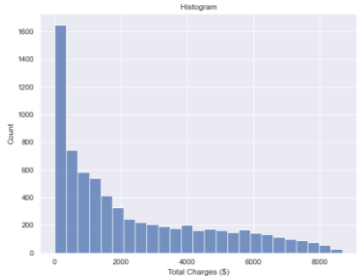
- Tenures (in years) are evenly spread except that over 1/7 of customers have tenures under 5 years and 1/7 of customers over 65 years



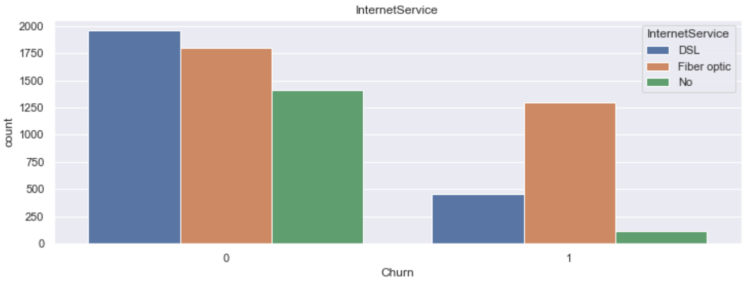
- Monthly Charges ($) are somewhat evenly spread except for a large spike of customers with charges $25 and under



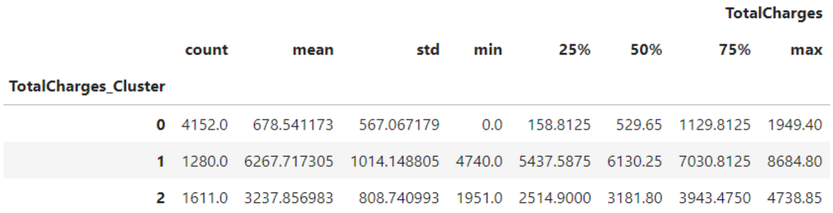
- Total Charges ($) is very skewed with a large spike at $500 and under, tapering off to greater values



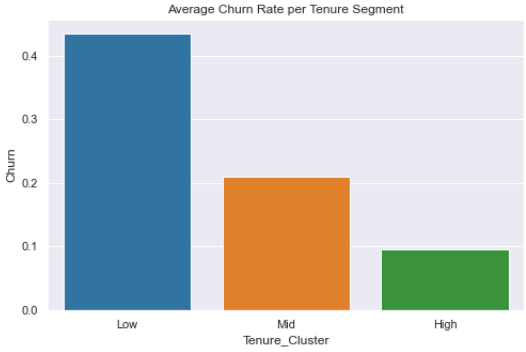
- Gender and Dependents are split 50/50
- Of the categorical features' options, the ones who churned most were:
  - Contract: Month-to-month
  - Online-Security: No
  - Tech-Support: No
  - Internet Service: Fiber optic (a surprise to me!)
  - Payment Method: Electronic check

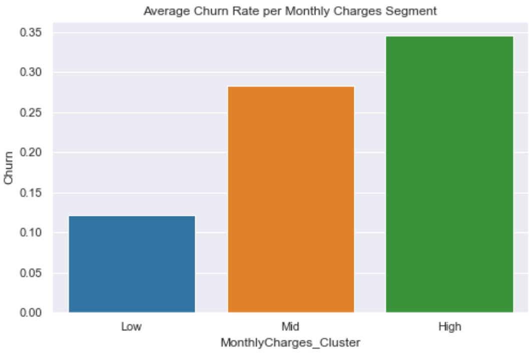


    - Online Backup: No
    - Device Protection: No
- Among the customer segments produced:
    - Low tenured customers churned most
    - High monthly charged customers churned most
    - Low total charged customers churned most

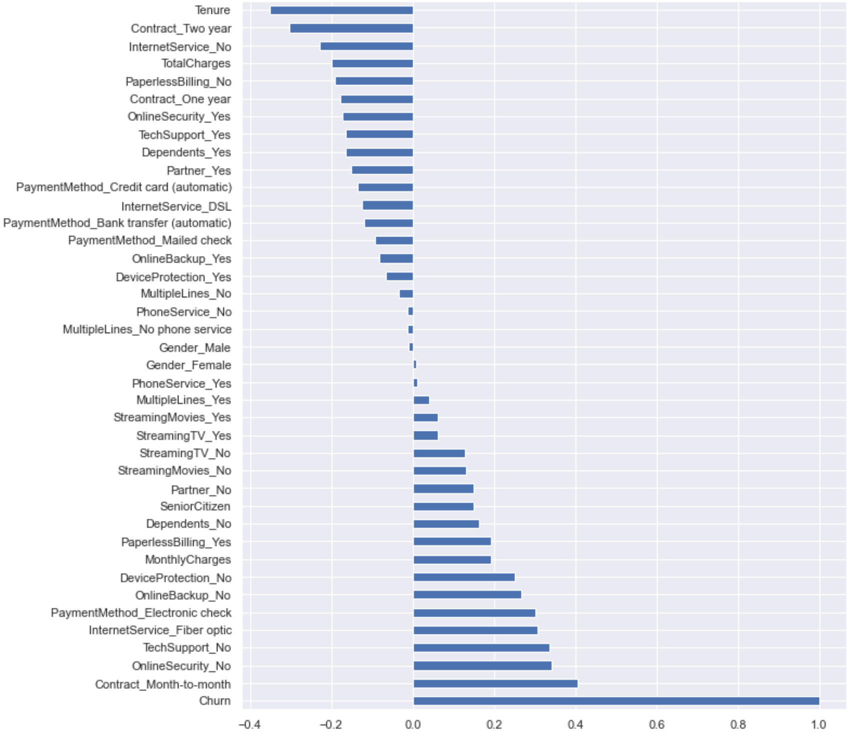


	TotalCharges							
	count	mean	std	min	25%	50%	75%	max
TotalCharges_Cluster								
0	4152.0	678.541173	567.067179	0.0	158.8125	529.65	1129.8125	1949.40
1	1280.0	6267.717305	1014.148805	4740.0	5437.5875	6130.25	7030.8125	8684.80
2	1611.0	3237.856983	808.740993	1951.0	2514.9000	3181.80	3943.4750	4738.85

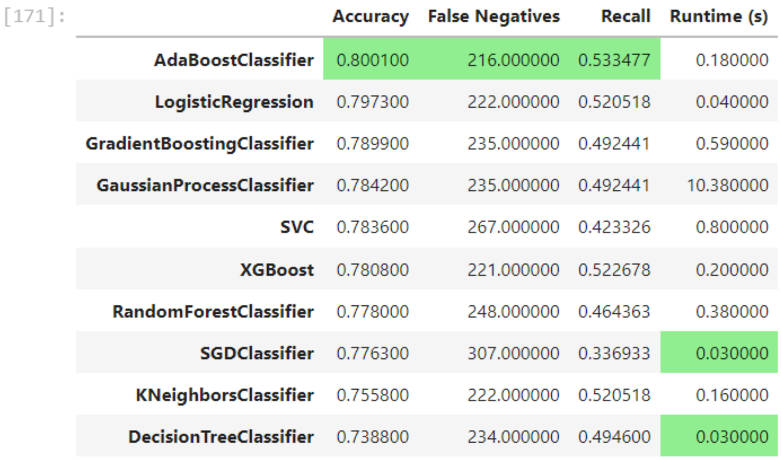




- Highly Correlating Features with Churn:
    - Tenure
    - Two-Year Contracts
    - No Internet Service
    - No Tech Support
    - No Online Security
    - Month to Month Contracts



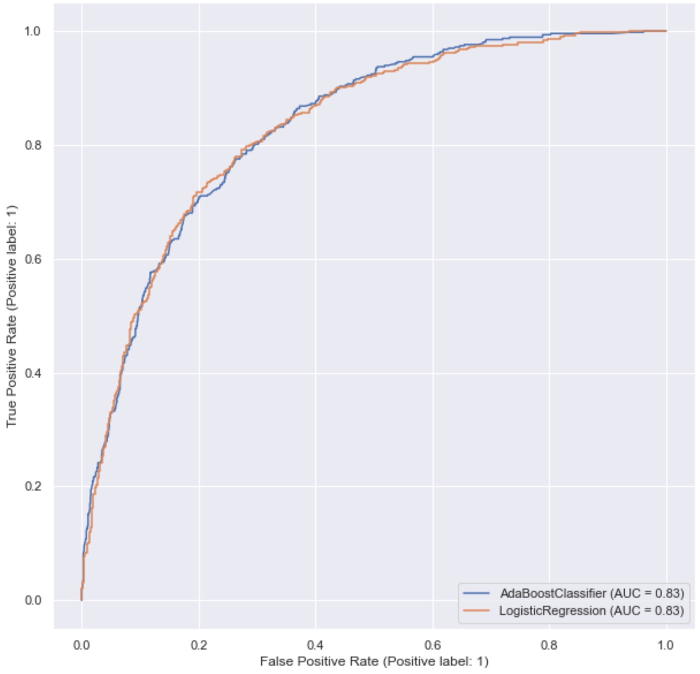
Now, modeling. In model exploration, where untuned models are compared to see raw performance on the dataset, almost all models performed the same, in regards to accuracy, at ¾ correct. The differentiator was recall score, where it could be shown which model can avoid false negatives. The best performers were AdaBoost and Logistic Regression.



[171]:

	Accuracy	False Negatives	Recall	Runtime (s)
AdaBoostClassifier	0.800100	216.000000	0.533477	0.180000
LogisticRegression	0.797300	222.000000	0.520518	0.040000
GradientBoostingClassifier	0.789900	235.000000	0.492441	0.590000
GaussianProcessClassifier	0.784200	235.000000	0.492441	10.380000
SVC	0.783600	267.000000	0.423326	0.800000
XGBoost	0.780800	221.000000	0.522678	0.200000
RandomForestClassifier	0.778000	248.000000	0.464363	0.380000
SGDClassifier	0.776300	307.000000	0.336933	0.030000
KNeighborsClassifier	0.755800	222.000000	0.520518	0.160000
DecisionTreeClassifier	0.738800	234.000000	0.494600	0.030000

AdaBoost and Logistic Regression were then hyperparameter tuned with GridSearchCV. Ultimately, AdaBoost's 0.555 recall score (1 is best, 0 is worst) performed better than Logistic Regression's 0.553. However, they both achieved an AUC score of ~0.83. Thus, AdaBoost was chosen as the final model.



The Final Results

Accuracy: 0.8001

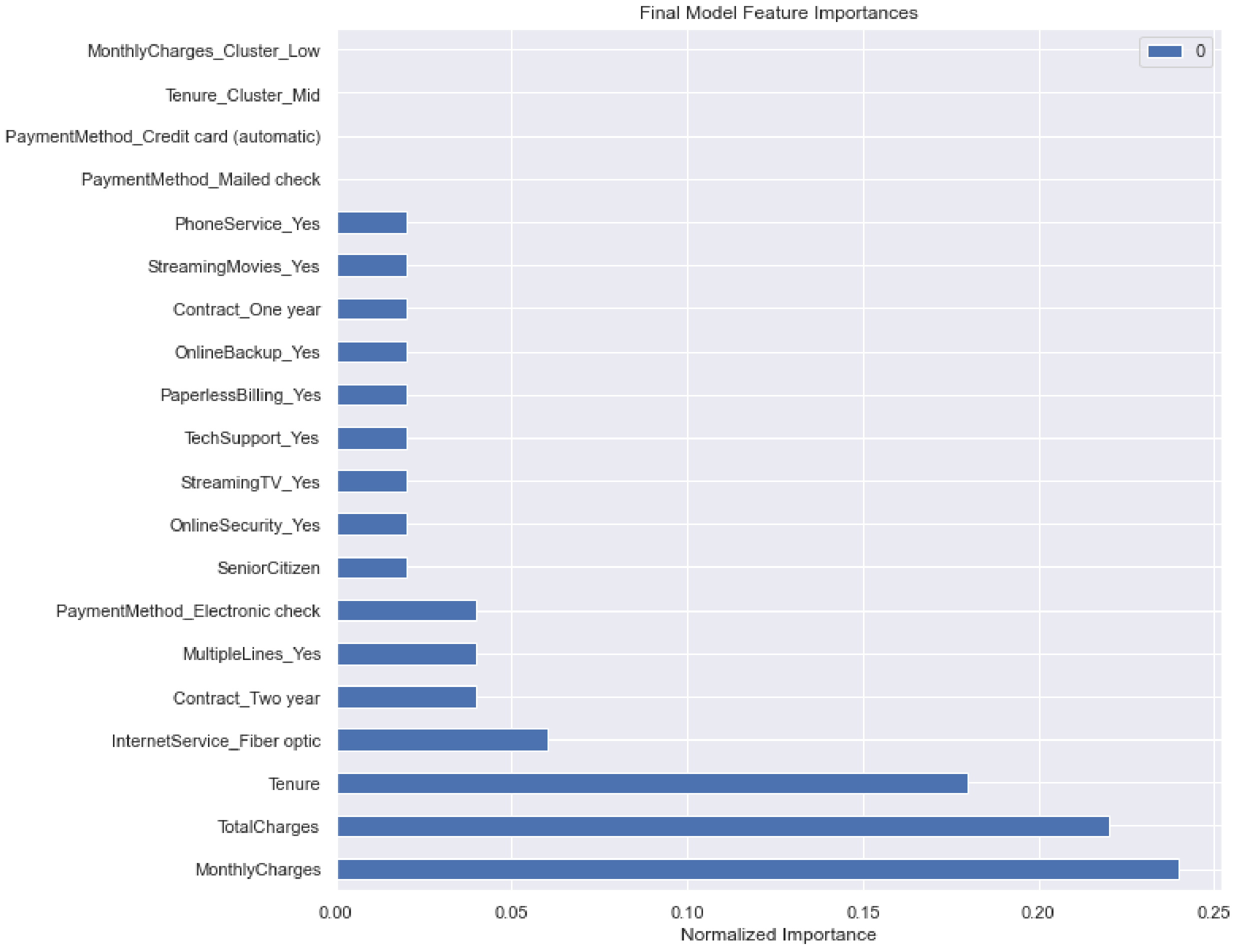
Recall: 0.5335

False Negatives: 216 / 1761 or 12.3%

AUC: 0.83

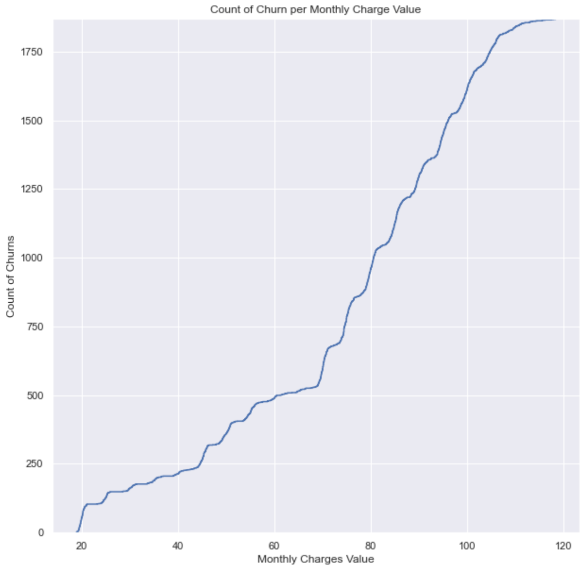
Interpreting the model

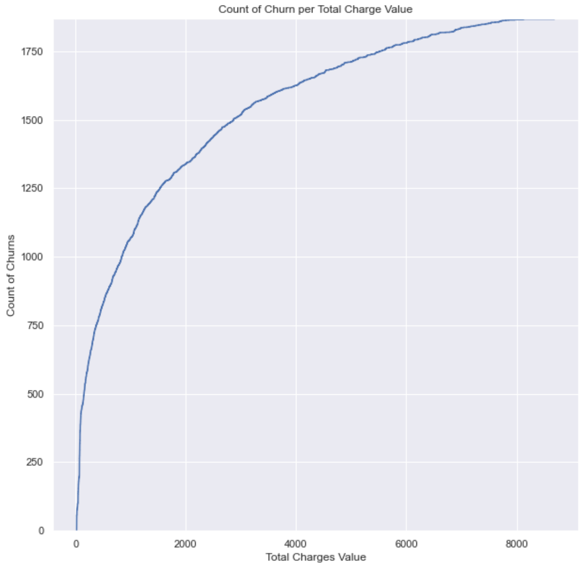
Monthly Charges, Total Charges, and Tenure make up about 65% of the predicting power for the final AdaBoost model.

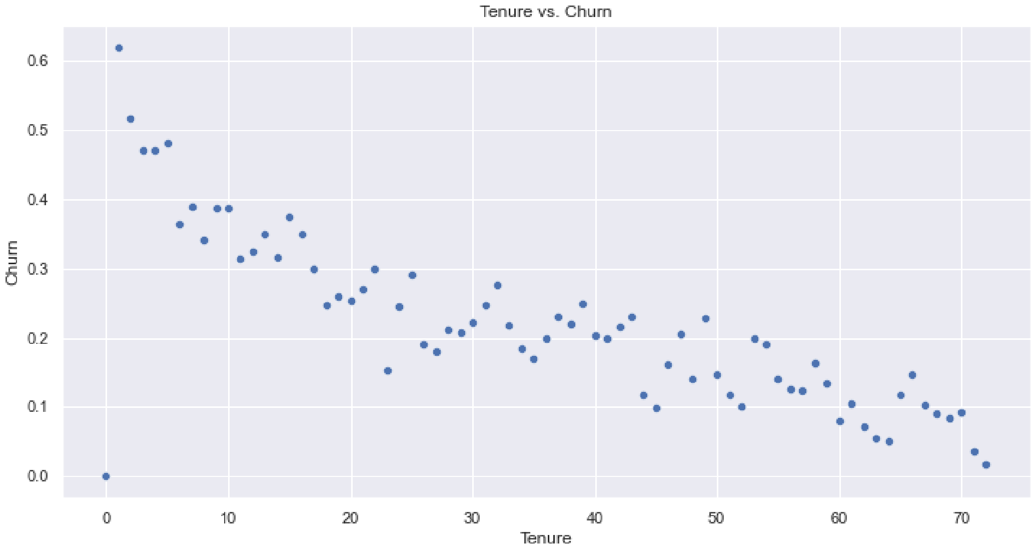


Focusing on the 4 most important features, and referencing to EDA in the iPython notebook…

1. *Monthly Charges:* Generally, the higher the charges the more likely to churn. Specifically, above $70 and the likelihood of churning sharply increases.
2. *Total Charges:* Customers with total charges under $1,000 were likely to churn, making up almost ½ of the total churns.
3. *Tenure:* The less tenure, the more likely to churn
4. *Internet Service – Fiber Optic:* Highest chance of churn in its category







How to Use the Findings

As the business, act on these predictions and try to retain the customers predicted to churn!

1. Reduce monthly charge rates
2. Incentivize customers with total costs below $1,000 / year to commit to more offerings
3. Focus on first 5 year customers especially, after that the likelihood to churn calms
4. Figure out what is happening with customers and fiber optic internet service!