Evaluating Popular Customer Churn Models

Christopher Ly
University of California, San Diego
La Jolla, CA 92093
chl818@ucsd.edu

Michelle Tran
University of California, San Diego
La Jolla, CA 92093
met016@ucsd.edu

Abstract

With the expansion of economic sectors and the rapid growth in competition, it is increasingly crucial for businesses to maintain a customer base. Identifying which customers are at risk of leaving can help marketing, sales, and engineering teams develop strategies to retain these customers. But with the exponential availability of data, information based models are more viable to predict the likelihood of customer's intention to switch providers, otherwise known in the field as churn. In this paper, we implement three traditional prediction techniques (Logistic Regression, Support Vector Machine, and AdaBoost Ensemble) as predictors based on new features to gain insight about what customers value the most. The experimental results show that SVM models are more effective compared to other models for customer churn prediction.

1 Introduction and Motivation

Over the last decade, the world entered the age of big data. And as more and more companies leverage this to their advantage in the market, it is important to have the most effective models to maintain a position ahead of competitors. One aspect of this is the ability to grow the customer base and maintain current customers from switching services to competitors, otherwise known as churning.

So the goal is clear, having measures in place to entice customers to stay and at the same time, not waste resources on customers who are unlikely to switch providers. Also being able to preemptively predict those most likely to churn to make efforts in retaining them. However, they're are different cases of churning: involuntary and voluntary cancellations [2]. The focus of this paper will be on voluntary cancellation as most companies do not typically consider involuntary cases as churning or abandonment.

2 Related Work

The problem of predicting a customer's likelihood to churn is not new. There has been a variety of machine learning techniques taken in approaching this problem. However, the effectiveness of these algorithms seems to depend on the data available to be used in training. Neural networks rely heavily on large amounts of data so with enough data and features available [4], they can prove effective in this task but may perform worse than traditional techniques [2].

Another approach is creating a new set of features which would then be used in predicting customer churn. With a new set of features presented in Huang's paper, they were able to achieve better prediction performance on all the machine learning techniques they used [3].

	Num of features	Num of datapoints	Num of not churned	Num of churned	% churned
Before ETL	19	7043	5174	1869	26.536987
After ETL	22	7032	5163	1869	26.578498

Table 1: Breakdown of the dataset

3 Dataset

3.1 Dataset Properties

The data we used to predict churn was from the telecommunications company, Telco, and after omitting customer ID, the data set contained 7,043 unique values for each of the following features per customer: gender, senior citizen status, partner status, if they have dependents, tenure in months, if they purchased phone service, if they have multiple phone lines, if they purchased internet service, if they have online security with their internet service, if they have online backup with their internet service, if they have device protection with their internet service, if they have tech support with their internet service, if they have TV streaming with their internet service, if they have movie streaming with their internet service, what type of contract they signed, if they have paperless billing, what type of payment method they use, monthly charges, total charges since being a customer, and if they churned. The total number of not churned customers and churned customers was 5174 and 1869, respectively (See Table 1).

We trimmed the data set to exclude all rows where customers had a tenure of 0 months, which meant they had just signed up for Telco and their total charges since becoming a customer was 0.

A general analysis for the churn rate by feature led us to a few observations. There appeared to be no distinction between genders for those who did and did not churn (See Figure 1). Most customers did not identify as senior citizens, and showed a slightly higher churn rate compared to senior citizens. Customers who were without a partner, or dependents were also more vulnerable to churning. Those who purchased phone service were likely to churn regardless of if they had multiple lines or not. Customers that bought Telco's internet services, and selected a fiber optic plan appeared to have a higher rate of churn compared to DSL internet services. For both the fiber optic and DSL plan, those who chose no online security, online backup, device protection, tech support, streaming TV, or streaming movies were less likely to continue their service. Month-to-month contracts displayed significantly higher rates of churn compare to one or two year contracts. Customers who opted for paperless billing showed higher rates of churn compared to those who did not. Those who paid by electronic check showcased the highest rate of churn relative to those who paid by mailed check, automatic bank transfer, or automatic credit card payment.

The distribution of the amount of customers by their tenure in months showed a heavy right skewed distribution for customers who did churn (See Figure 2). This is expected based on the high churn rate of customers who had month-to-month contracts. Customers who did not churn exhibited a bimodal distribution for their tenure. Plotting the number of monthly charges for customers who did churn appeared to be slightly left skewed, while those who did not churn showed a heavy right skewed distribution. The distribution for total charges for all customers, regardless of churn or not, indicated a right skewed distribution.

There also was a high correlation of 0.826 between total charges and tenure (See Figure 3).

4 Methods

Before training the models, our findings from exploring the dataset suggested that the data needed a bit of cleaning. To prepare the features, we observed that customers have the option to buy additional services such as online security, online backup, device protection, tech support, streaming TV, or streaming movies when purchasing internet services. We created a feature

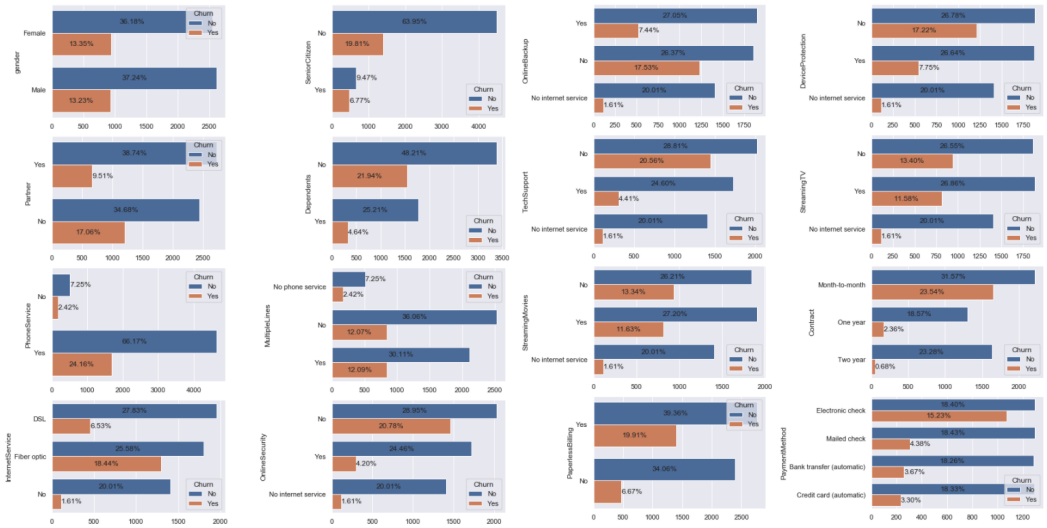

Figure 1: Churn rate/occurrence per feature

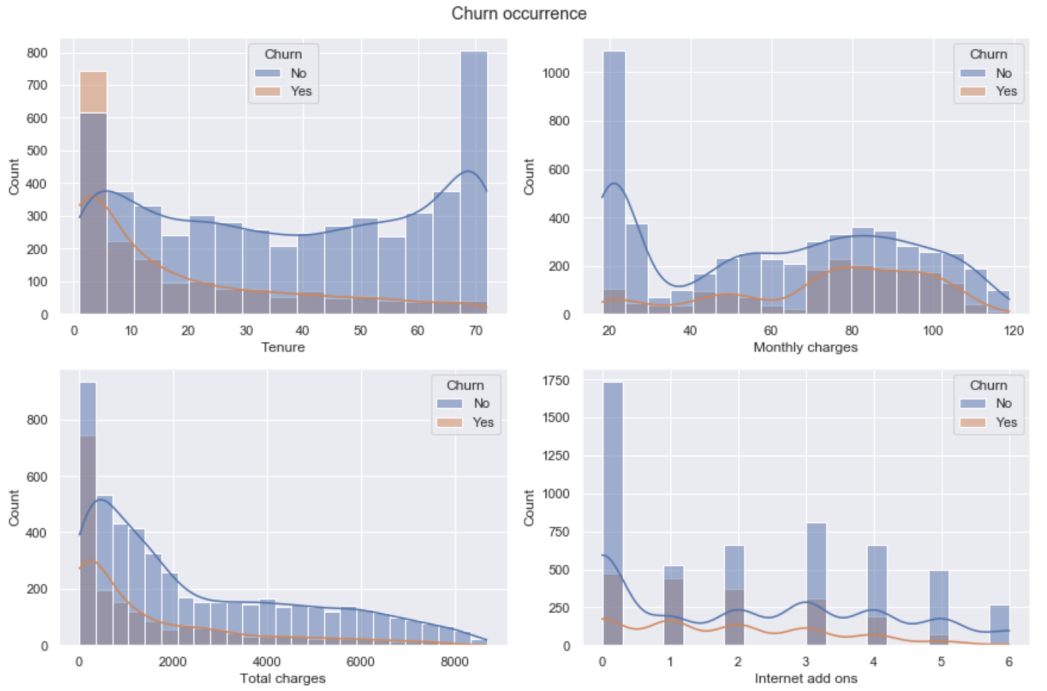

Figure 2: Churn rate/occurrence per feature

	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	PaperlessBilling	MonthlyCharges	TotalCharges	InternetAddOns
gender	1.000	0.002	0.001	-0.010	-0.005	0.008	0.012	0.014	-0.000	0.014
SeniorCitizen	0.002	1.000	0.017	-0.211	0.016	0.008	0.156	0.220	0.102	0.068
Partner	0.001	0.017	1.000	0.452	0.382	0.018	-0.014	0.098	0.319	0.205
Dependents	-0.010	-0.211	0.452	1.000	0.163	-0.001	-0.110	-0.112	0.065	0.030
tenure	-0.005	0.016	0.382	0.163	1.000	0.008	0.005	0.247	0.826	0.495
PhoneService	0.008	0.008	0.018	-0.001	0.008	1.000	0.017	0.248	0.113	-0.092
PaperlessBilling	0.012	0.156	-0.014	-0.110	0.005	0.017	1.000	0.352	0.158	0.183
MonthlyCharges	0.014	0.220	0.098	-0.112	0.247	0.248	0.352	1.000	0.651	0.725
TotalCharges	-0.000	0.102	0.319	0.065	0.826	0.113	0.158	0.651	1.000	0.746
InternetAddOns	0.014	0.068	0.205	0.030	0.495	-0.092	0.183	0.725	0.746	1.000

Figure 3: Correlation between features

called InternetAddOns that sums the total number of services a customer purchased in addition to their internet service. For customers that did not purchase internet service or did but with no additional services, the value for InternetAddOns was considered to be 0.

Based on the high correlation previously mentioned between total charges and tenure, we omitted total charges to prevent interference in model interpretation due to the nature of correlation versus causation.

The dataset needed cleaning in terms of the values as well. Since the range of values between the features varied greatly, we applied a standard scaling which can be observed in the following equation:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where

μ is the mean of x

σ is the standard deviation of x

After applying one hot encoding to the categorical features, we observed that a majority of the datapoints were retained and ended up with more features than before.

As the data has now been prepped for training, the training and test dataset were created using a 80/20 split where 20% of the data is reserved for testing.

4.1 Logistic Regression

Logistic Regression is a linear machine learning model that uses the logistic function for activation. The model provides a probability between class 1 and class 2 for binary classification. Using sklearn's implementation, we compare the default parameters, aside from setting the solver as liblinear, to the best ones found via a grid search.

4.2 Support Vector Machine

Support Vector Machine, or SVM for short, is another type linear machine learning model that finds "support vectors" to help determine the optimal decision boundary that separates the class with the highest margin, distance between the class. However, SVM is not strictly a linear separator since it depends on the kernel used. Using sklearn's implementation for Support Vector Classification, we compare the default parameters, aside from setting the gamma as auto, to the best ones found via grid search.

4.3 AdaBoost

AdaBoost is an ensemble based machine learning model that combines the prediction of multiple weak classifiers to create an overall prediction [1]. Due to this ensemble approach, AdaBoost

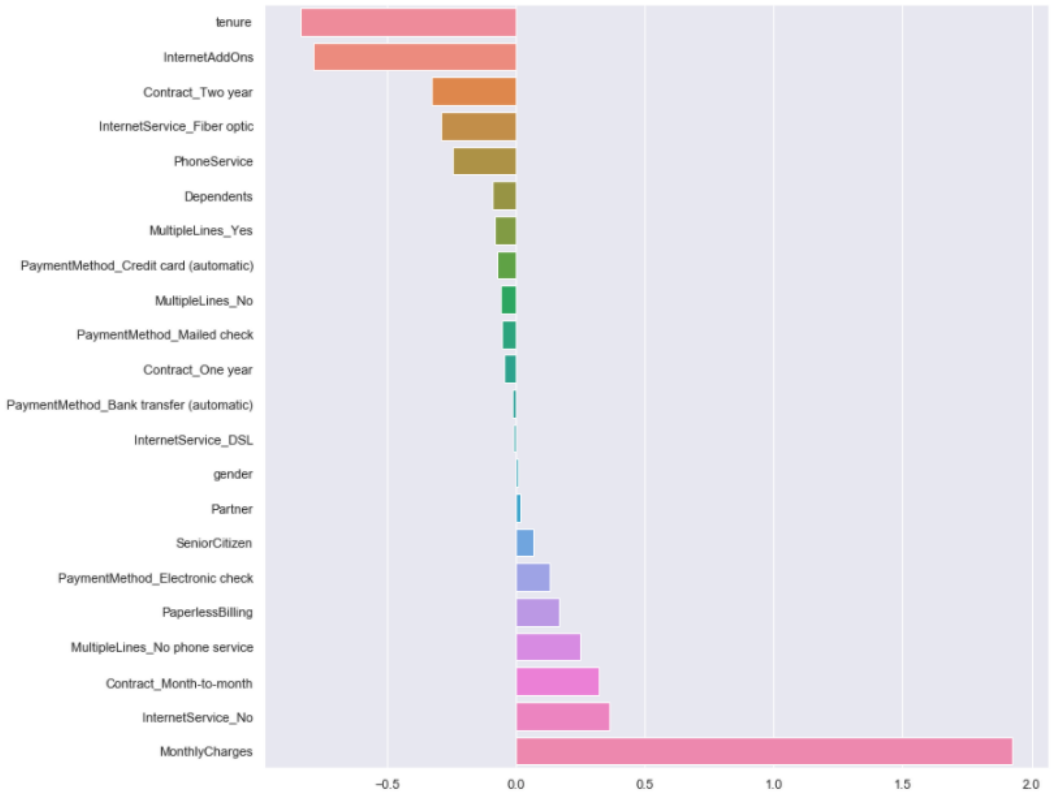

Figure 4: Weights of Logistic Regression

is a nonlinear machine learning model. Using sklearn's implementation, we compare the default parameters to the best ones found via a grid search.

5 Results

To evaluate the effectiveness of the trained models, we used 3 metrics: accuracy, f1 score, and f_β score with a β value of 5. From Table 2, we can see that the accuracy from the default parameters to the optimal ones found through grid search were actually lower. We chose a larger value for β as wanted to mitigate the amount of false negatives from the models.

After running grid search on the logistic regression model using 5 fold cross validation, it determined 'C' to be 0.001 with a maximum number of iterations of 100, L_2 penalty, and a tolerance of 0.0001. This gave an accuracy of 74.55%. After running grid search on the SVM model using 5 fold cross validation, it determined 'C' to be 0.01 with gamma set to auto, a linear kernel, and the maximum number of iterations to be 1000. This gave an accuracy of 74.83%. The resulting parameters from grid search on the AdaBoost model was a learning rate of 1 and 100 estimators to achieve an accuracy of 71.99%.

6 Discussion

Throughout this experiment, we trained several models with the task of predicting whether or not a customer would be likely to switch services to another provider. The resulting weights of the logistic regression model are in Figure 4. This led us to the conclusion that having a longer tenure and more additional internet services influences customers to not churn. Customers that utilize more Telco services might find it inconvenient to switch all their services to another competitor. However, customers with high monthly charges are much more likely to churn. From the AdaBoost model, we can conclude that tenure and monthly charges are the most important

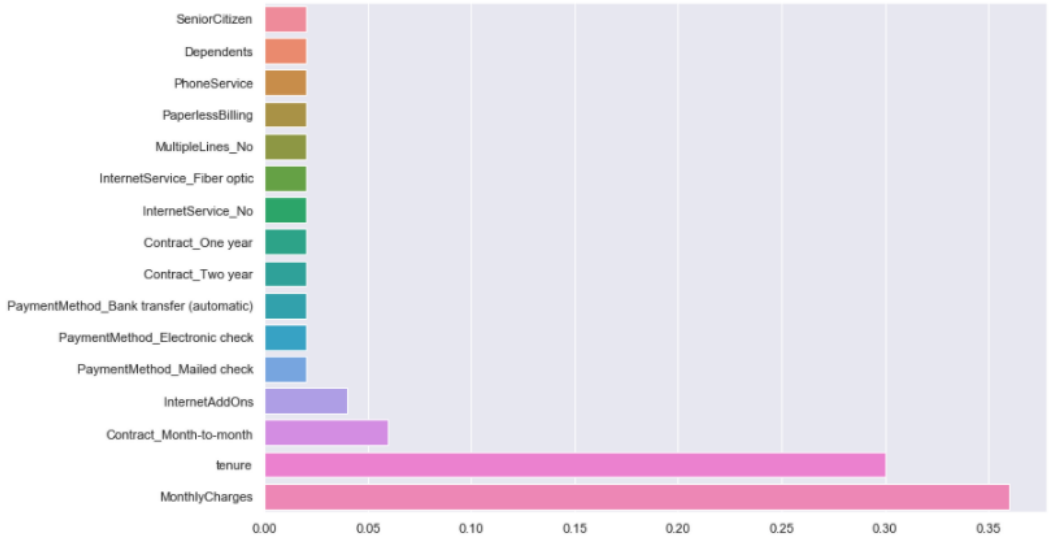

Figure 5: Feature importance summary of AdaBoost

factors for predicting churn (See Figure 5). Contradictory to our logistic regression model, the AdaBoost model considers additional internet services less significant than tenure and monthly charges.

Across all the models, our low F1 scores indicate that despite having decent accuracy, they are not effective in capturing who is at risk of churning. This is better observed from the F_β scores for Logistic Regression and AdaBoost models. Since we consider a β of 5, we heavily weigh importance on recall, i.e. minimizing the number of false negatives. Because of the low F_β scores in the context of high accuracies, this means that the models are ineffective at detecting churners. This is due to accuracy being a misleading evaluation metric for an imbalanced dataset such as this one. The models can achieve high accuracy by classifying people under the more abundant class but taking the F scores into account tells a different story.

While Logistic Regression and AdaBoost perform well in terms of accuracy, as we've discussed how these models were ineffective in the task of predicting those most likely to churn. Although SVM achieved a low accuracy, it was the only model to achieve a well performing F_β score. This means SVM is actually a better model for the specified task. This is because it better captures who is likely to churn by sacrificing the amount of false positives in exchange for minimizing false negatives. In terms of the market, while more resources are spent trying to retain people unlikely to churn, using SVM means that we would be retaining more of those who would.

The largest issue was our lack of domain knowledge as it relates to telecommunication. Because of this, it was difficult to use prior knowledge on how to best reconfigure the features such as how each of them relate and impact another in our attempt to optimize the prediction models. An improvement would be spending additional time to thoroughly research information into the domain and have a holistic grasp as to how to use the data we obtained and prepare it for model training. Another potential improvement would be obtaining more data but this might prove to be difficult as churn data is not easily accessible, especially data that has similar features or even come from the same market. Having more data would provide the models a diverse and representative sample of the data generating process. This is merely a limitation of the public domain as companies lose their competitive advantage by providing such data freely. An extension from here would be to explore deeper into the balance of precision and recall summarized in the F_β scores. This could lead to more effective models, those achieving a better balance of F_β score and accuracy, which could be implemented in real business decisions.

	Accuracy before grid search	Accuracy after	F1 Score	F_β
Logistic Regression	80.028429	77.256574	0.577444	0.547789
SVM	79.459844	56.289979	0.557656	0.806859
AdaBoost	79.175551	78.678038	0.551302	0.514229

Table 2: Evaluation metric results

References

[1] Sanjoy Dasgupta. Boosting lecture notes, 2013.

[2] David L. García, Àngela Nebot, and Alfredo Vellido. Intelligent data analysis approaches to churn as a business problem: a survey. *Knowledge and Information Systems*, 51(3):719–774, 2016.

[3] Bingquan Huang, Mohand Tahar Kechadi, and Brian Buckley. Customer churn prediction in telecommunications. *Expert Systems with Applications*, 39(1):1414 – 1425, 2012.

[4] Shin-Yuan Hung, David C. Yen, and Hsiu-Yu Wang. Applying data mining to telecom churn management. *Expert Systems with Applications*, 31(3):515 – 524, 2006.