Analysing Customer Churn in the Telecom Sector

Machine Learning Group Assignment

Submitted by: Huy Nguyen, Denton Li, Shreya Verma

Contents

1. Background and Motivation

The telecom sector is made up of companies that make communication possible on a global scale, whether it is through the phone or the Internet, through airwaves or cables, through wires or wirelessly. However these days, customer retention has become a major issue in the world of phone and internet services. Customers not only leave the telecom company but also make it a point to spread it in their social circle. Thus this makes it necessary to look at this problem of customer churn for the telecom companies, and address it by customer retention. This serves as the motivation for working on the Telecom Sector dataset and solving this business problem.
Since in this industry, serving one more customer adds only subtle, if any, variable costs to the company, keeping a customer would mean a certain future inflow of profits.
Thus, it is definitely worth it to invest in keeping the customers stay, even if it means giving promotions or gifts. However, it might be costly or revenue hurting to promote to every customer just to hope that those wanting to churn would stay. Therefore, an ability to identify customers that are likely to churn is valuable.
We are doing research for a (hypothetical) telecom company operating in the USA, Better Tel, to better improve its profits. More specific background information is listed in section 1.1 and 1.2 below.

1.1 What is customer churn?

Customer churn refers to the customer attrition or loss of clientele. In other words, It is most commonly expressed as the percentage of service subscribers who discontinue their subscriptions within a given time period.

1.2 Consequences of high churn rate?

Effects on Revenues. Churn rate can affect the company's revenues adversely. Research by Bain & Co shows that increasing customer retention rates by just 5% can increase profits by somewhere between 25% and 95%. So with such a dramatic change in the revenues reducing churn rate becomes an extremely imperative step in the whole profit maximization process.
Retaining bigger customers would generally cost less than acquiring new customers. That being said, customer loyalty and excellent customer service go hand in hand. Thus telecom companies nowadays focus more on these and leverage their core competencies to generate higher returns and rely on customer retention for greater profits.

2. Business and Statistical questions

There are two business questions we are interested in to improve the profits of Better Tel. The first one is what features of customers that decide whether the customer would churn. Knowing this would help the company better serve and retain the customers or target the right customers with demographic traits that make the customer unlikely to churn.
The second business question is whether the company can have a better way to identify the customers that would churn than just assuming all of them would churn and put the effort in every single customer to prevent it, or than just assuming that all of them would not churn and do nothing.
A statistical translation of the first business question is what explanatory variables determine whether a customer would churn, and what are the directions and strength for those variables?

A statistical translation of the second business question would be to build a machine learning algorithm that can predict with reasonable accuracy, with a high true-positive rate and a low false-positive rate, whether a customer would churn.
Finding the answers for these questions is important for several reasons:

1. It improves efficiency in allocating the company's cost on promotions. Customers that are highly likely to renew their contracts do not need extra incentives. Instead, incentives should be given to customers with the probability of leaving around or above 0.5. In other words, these customers are on the fence of leaving and continuing with the telecom services. Therefore, incentives should be given to these groups.
2. By knowing the important features that determine the churn ratio, we can further promote that service to new customers so that the retention rate will be higher in the future. In addition, services that contribute to a high churn rate will be investigated to find out the reason why it does not motivate customers to stay.
3. Based on the demographics of customers with higher probability of staying, we can target the new customers that have similar demographics to improve the churn rate.

Regarding the scope of the statistical question, we do not determine the cost and benefits of the investment to keep a customer. In reality, with enough time and information, we would try to find out how much revenue would be generated by preventing a churn. Also, we do not predict how long a customer would stay with the company. Finally, the adjustment for the threshold of probability for a customer to churn, or for the sensitivity is not considered in this model, as this adjustment requires cost-benefit-analysis. However, the final model we provide would be able to give a probability for an observation to churn, so the company would utilize it and set a reasonable threshold as needed.

3. Data set:

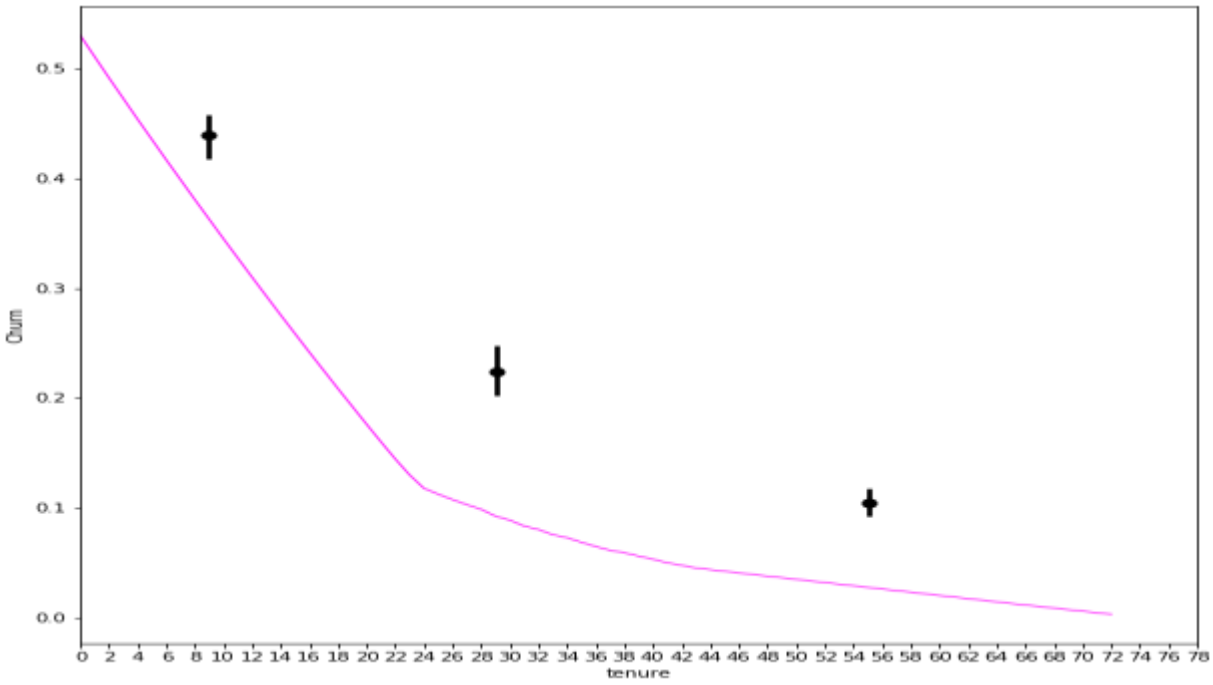
3.1 Description

The data set about "Telco Customer Churn" on Kaggle includes the information about the services that customers used, their account information, demographics and whether churn happened within last month. In details, the data set includes the information of 7043 customers with 21 columns (see appendix 1 for details of the variables). We assume that the data is sampled all over the USA.

3.2 Steps for Data Cleaning

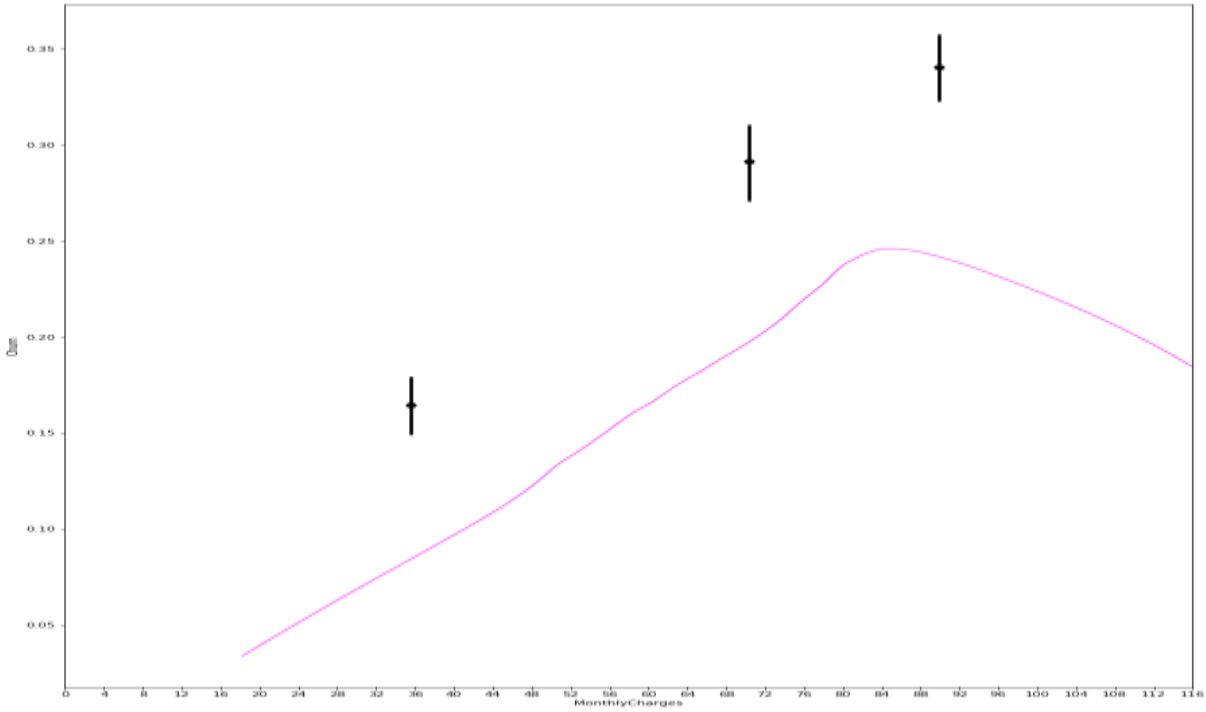
1. The data is in csv format and is imported to the jupyter notebook, and irrelevant columns are dropped. There are no missing values.
2. Now we transform the categorical variables, and we create a new column-"Total Charges" to obtain numeric values to gauge the impact of tenure and monthly charges.
3. We have 14 binary variables so we transform them to add 0 and 1 to different categories.
4. So, for the variable gender we assign Female as 0 and Male as 1, for internet service we replace it with two columns "DSL" and "Fibre Optic", with "no internet service" as the default, for contract we add two more columns "one year" and "two year", with "one month" as the default. Lastly for payment methods 'electronic cheque' is set as default and 3 other columns- 'credit card', 'bank transfer', 'payment method' and 'mail'.
5. Finally we deleted these original columns 'payment method', 'Internet service', and 'contract'.
6. Now we check for outliers, in our dataset outliers can only occur in Tenure, MonthlyCharges, and TotalCharges, since other variables are categorical.

7. However, after we plotted the lowess lines for those three variables, we found that some other variables should be created.
   The lowess line for Tenure:



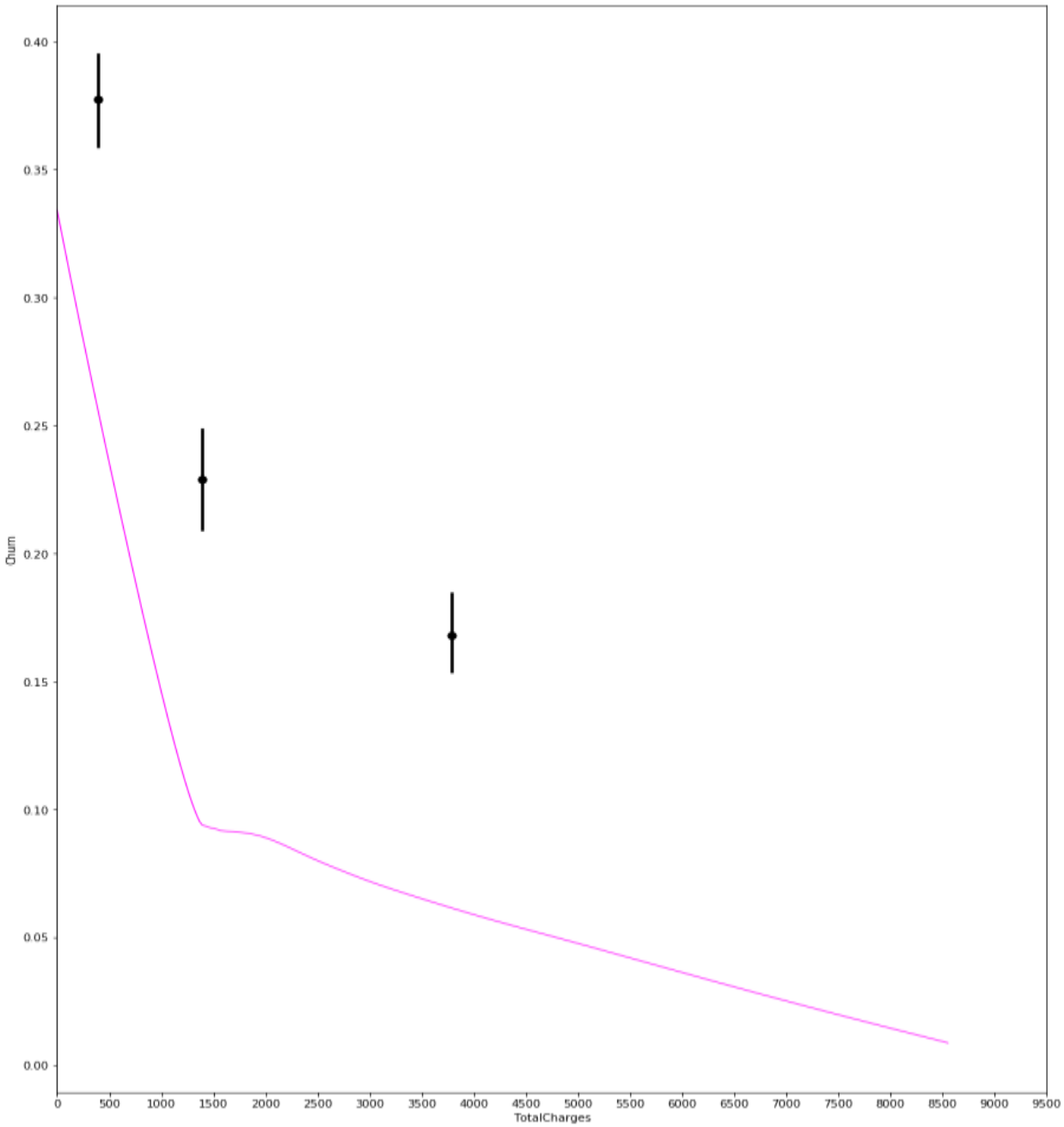
From this graph, we can see that the slope changed after the tenure goes above 24, so we added another categorical variable: 'tenure over 24', which has the value of 1 if the value of tenure is above 24, and 0 otherwise. Next, we created an interaction term between “tenure” and “tenure over 24” to better capture the change in the slope.

Then we proceeded to check the lowess line for the MonthlyCharges:



From the graph, we found out that after the value of monthly charges reaches over 84, the slope changes. So we did the same thing and created two more variables.

Then we went on checking the lowess line for the TotalCharges:

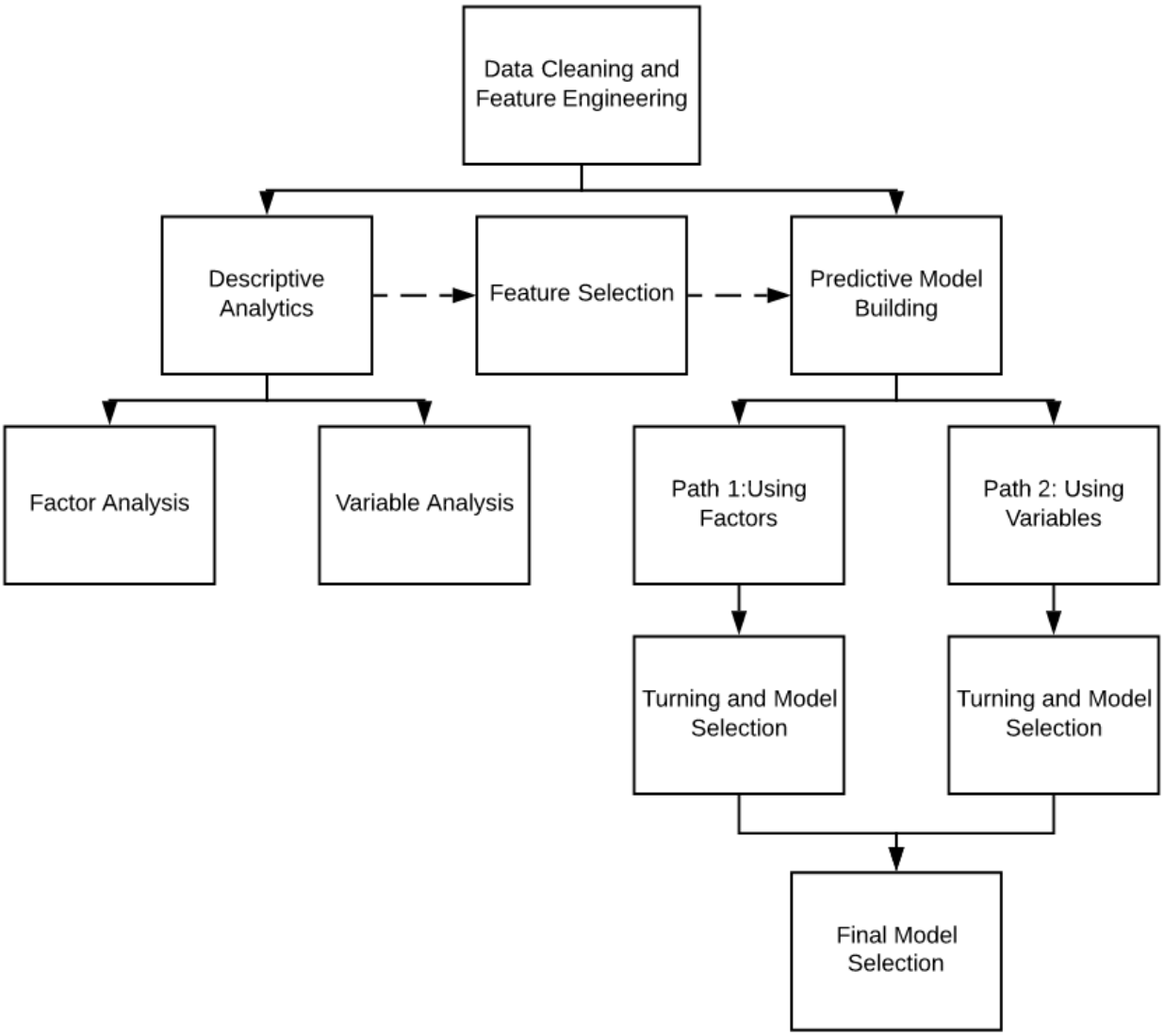


The result shows that the slope changes after the TotalCharges reaches 1500, so we added another two variables.

8. Then we referred to the correlations between each variable and "Churn", dropping those with the coefficient below the absolute value of 0.1.
9. After the analysis, we will proceed with 21 columns and 7043 observations.

4. Discussion about our supervised learning workflow and models

The following flowchart describes the workflow that we follow to come to our final model and conclusion.



4.1 Descriptive machine learning analysis

4.1.1 Principle Component Analysis

Principle component analysis (PCA) is used to find out the most relevant factors for our analysis. In order to determine if the data set is suitable for PCA, We use the Kaiser-Meyer-Olkin (KMO) Test and Barlett Test of sphericity. Since the p-value of the Barlett Test is 0 and the KMO value is 0.618, we can do PCA on this data set.

4.1.2 Factor description

According to the factor loadings table (Exhibit 2), the 6 factors can be described as follow:
Factor 1, Dedication and loyalty: This factor measures how much phone service tenure term and money a customer invests in the company.

Factor 2, Frugal-stickerness: This factor captures a trait of observations that stick with the company for a long time but spend little on services.
Factor 3, DSL loneliness: This factor exhibits a trait of subscribing DSL service, having no dependants, and needing tech/security support.
Factor 4, Credit card lover: This factor showing that one prefers to pay by credit card.
Factor 5, Family provider: This factor shows one tends to have dependents and partners.
Factor 6, Extreme terms signers: This factor captures a trait that prone to signs either a one-month contract or a 2-year contract.

1. Coefficients of the factors

We want to determine the coefficient of the above factors by creating a regression model. Since the outcome 'churn' is binary, we find that logistic regression is appropriate. In order to determine the best model, we use RandomizedSearchCV with different hyperparameters ('classifier_C', 'classifier_solver') in the pipeline.

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
0	-0.458	-1.129622	0.166936	-0.049821	0.12834	-0.008645

According to the best model, factor 2(with the coefficient of -1.13) is the most important, followed by factor 1(with the coefficient of -0.46). The other factors do not have a high correlation with the churn rate.

2. Coefficients of the features

We also want to determine how important the original features are in determining the 'churn' rate. In order to find the best model, we use a randomized search that tries two algorithms "Random Forest Classifier" (with hyperparameters "max depth" and "n-estimators") and "Logistic Regression" (with hyperparameters "C" and "solver")
The best model turns out to be logistic regression with an error rate of 0.187. Below is the coefficients of the features.

	SeniorCitizen	Partner	Dependents	tenure	OnlineSecurity	TechSupport	PaperlessBilling	MonthlyCharges	TotalCharges	DSL
0	0.091742	0.014146	-0.077732	-2.605013	-0.2467	-0.178881	0.174144	-0.173058	0.969039	0.633721

Fiber optic	oneY_contr	twoY_contr	CreCard	Banktran	tenure over 24	TO24INT	month over 84	MON84INT	total over 1500	TOCH1500
1.049655	-0.288498	-0.669438	-0.098065	-0.070221	-0.950066	2.697106	-1.562068	1.899466	0.507873	-1.326188

From the result, we can see that the most important variables are tenure and monthly charges. When tenure is less than 24, every increment has a significant effect in lowering the chance of churning. After the monthly charge is over 84, increment in every dollar increases the chance for churning. In addition, subscribing for internet service (DSL or Fiber optic) increases the chance of churning.

4.2 Predictive Machine Learning Analysis

We consider two paths in building the predictive algorithm based on the uses of either features or the factors created in the previous part.

Path 1: Building an algorithm using 21 features

Feature selection:

According to the previous step, all features are not equally important. To simplify the model, we want to keep only the more important features and drop the unimportant features. We consider two cut-off points of the absolute values of the logistics regression coefficients, 0.1 and 0.25. The reason why we used two cut-off points is to give us more selection on building a good model.
With 0.1 as cut off point, the dropped features are 'SeniorCitizen',' Partner', 'Dependents', 'CreCard', 'Banktran'.
With 0.25 as cut off point, the dropped features are 'SeniorCitizen',' Partner', 'Dependents', 'CreCard', 'Banktran', 'TechSupport', 'PaperlessBilling', 'MonthlyCharges', and 'OnlineSecurity
In order to find the best set of features for the final model, we try different models with both sets of features. Below are the error rates for each model and cut-off point..

Cut-off point	Logistic Regression	Random forest
0.1	0.183	0.192
0.25	0.194	0.197

Since cut-off point 0.1 performs much better than the cut-off point 0.25 in those two algorithms, we think that it will perform better in nearly every algorithm and decide to choose 0.1 as our cut-off point.

Model selection

With the subset of features that have coefficients > 1, we try different algorithms in addition to Logistic Regression and Random Forest and compare their error rates.

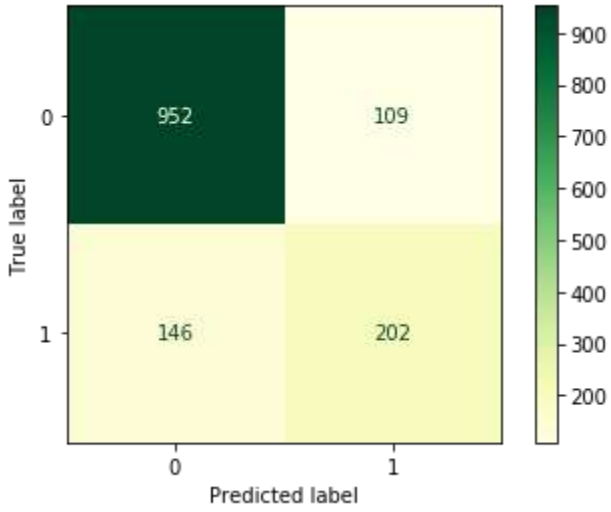
XGBoost-Tree	LGBM	Support Vector	KNN	Voting classifier*	Naive Bayes	Stacking
0.192	0.185	0.194	0.711	0.182	0.398	0.183

Path 2: Building an algorithm using 8 factors

Since we use only 8 factors from the Principle Component Analysis, there is no need to perform feature selection in this path. Similar to path 1, we try different models and compare their error rates:

Logistic Regression	Random forest	XGBoost-Tree	LGBM	Support Vector	KNN	Voting classifier	Stacking
0.194	0.181	0.186	0.184	0.202	0.704	0.184	181

Since the Stacking model in path 2 gives the best performance, our final predictive model would be: First transforming the data using the same PCA fitted function selecting 8 components and then using the Stacking model to predict it. The accuracy rate given the test set is around 81.9%, and given the whole dataset to train, the accuracy rate is expected to improve. Besides its accuracy, another merit for this model is that it has a high true-positive rate compared to its false-positive rate.



Comparing the final model to the null model

Null Model 1: Assuming that all customers would churn.

Although this null model has a 100% true positive rate, it also has a 100% false positive rate. The accuracy of the Null Model 1 would be 24.7%, and customers who would not churn, accounting for 75.3% of the total customers, would be given incentives to stay, which is a waste for the company's resource.

Although our final model has a true positive rate of only 58%, the model has a low false positive rate of only 10.3%, only wasting its promotion to 7.7% of the total customers. In addition, the threshold of the prediction to be changed to improve the true positive rate at the expense of the false positive rate.

Null Model 2: Assuming that all customers would not churn.

This model has an accuracy of 75.3%. However, it does nothing on keeping the customers. Our final model, though would generate a few false alarms, would identify 58% of customers that would churn.

Thus, our final model outperforms these two null models.

5. Communication of Results

Managerial Insights:

To answer the first business question, the features that affect the customers' churning, we will illustrate the most important factors or features.

We concluded two most important factorss that affect the customers' churns:

1. ***Dedication and loyalty: This factor measures how much phone service tenure term and money a customer invests in the company.***
2. ***Frugal-stickerness: This factor captures a trait of observations that sticks with the company for a long time but spend little on services.***¶

These two factors are both negatively correlated to churning, meaning that the more a customer has any of these two, the less likley he/she will churn.

Then how can these two factors guide the actions of the company? Factor 1 shows that telecom companies should target affluent customers and provide good services to them, maybe by allocating more

customer service teams to them. Because if people invest a lot in money, and if good service makes them stay, then they are unlikely to churn, and the affluent customers can afford to invest monetarily. Factor 2 shows that people who spend a little could be drawn by promotions or discounts to made stay. Better Tel can provide low-end service to attract “savers”, who if given the incentive to sign long contracts, are not likely to churn.
From the variable analysis, the most important features are tenures and monthly charges related features and internet service.
We can conclude that Better Tel should make an effort to make customers stick with it for more than 24 months. However, it should be careful about launching any services more expensive than 84 dollars per month. Finally, the company should do a customer survey to determine why those who subscribed to internet service(DSL or Fiber optic) tend more to churn.

For the second question, we built a predictive machine learning model to help the company identify the churning customers.
In order to find out the best model for predicting whether a customer would churn we run multiple machine learning algorithms, and after rigorous comparisons, we come up with a model that has an accuracy rate of 81.9%. Another insight that one gains from this model is that the number of accurate predictions for predicting the customer that would actually churn is really high which gives the Better Tel an opportunity to only reach out to the right customer subgroup who can potentially turnover to a competitor service provider. Hence that saves promotion costs as well as services, and the company would not advertise or try to retain a customer who is already loyal to the Telco.
Finally, the model can predict a probability for a customer to churn. If the company wants to retain a higher proportion than 58% of the customers that would churn, it can lower the threshold(now it labels a customer to be churning if he/she has a probability of churning larger than 50%) in the labeling. After conducting a cost-benefit analysis, the company should be able to determine an optimal number of percentages.

Appendix

List of variables in the data set:

- Customer ID
- Gender: Whether the customer is a male or a female
- SeniorCitizen: Whether the customer is a senior citizen or not (1, 0)
- Partner: Whether the customer has a partner or not (Yes, No)
- Dependents: Whether the customer has dependents or not (Yes, No)
- Tenure: Number of months the customer has stayed with the company
- PhoneService: Whether the customer has a phone service or not (Yes, No)
- MultipleLines: Whether the customer has multiple lines or not (Yes, No, No phone service)
- InternetService: Customer's internet service provider (DSL, Fiber optic, No)
- OnlineSecurity: Whether the customer has online security or not (Yes, No, No internet service)
- OnlineBackup: Whether the customer has online backup or not (Yes, No, No internet service)
- DeviceProtection: Whether the customer has device protection or not (Yes, No, No internet service)
- TechSupport: Whether the customer has tech support or not (Yes, No, No internet service)
- StreamingTV: Whether the customer has streaming TV or not (Yes, No, No internet service)
- StreamingMovies: Whether the customer has streaming movies or not (Yes, No, No internet service)
- Contract: The contract term of the customer (Month-to-month, One year, Two year)
- PaperlessBilling: Whether the customer has paperless billing or not (Yes, No)
- PaymentMethod: The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
- MonthlyCharges: The amount charged to the customer monthly
- TotalCharges: The total amount charged to the customer
- Churn: Whether the customer churned or not (Yes or No)

  (Source Kaggle.com)

  (Below are columns created)
- Tenure over 24: A categorical variable indicating whether the tenure is over 24 months (1: tenure > 24, 0: tenure <= 24).
- TO24INT: The interaction term(multiplication) between "tenure over 24" and "tenure".
- Month over 84: A categorical variable indicating whether the monthly charge is over 84 (1: MonthlyCharges > 84, 0:MonthlyCharges <= 84).
- MON84INT: The interaction term(multiplication) between "MonthlyCharges" and "Month over 84".
- Total over 1500: A categorical variable indicating whether the total charge is over 1500 (1: TotalCharges > 1500, 0: TotalCharges <= 1500).
- TOCH1500: The interaction term(multiplication) between "TotalCharges" and "Total over 1500".

Factor Loadings

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
SeniorCitizen	0.119515	-0.241371	0.051698	-0.005843	-0.124223	-0.024480
Partner	0.336107	0.229366	-0.232347	-0.044593	0.258328	-0.042005
Dependents	0.087824	0.375421	-0.477119	-0.059987	0.778721	-0.100569
tenure	0.845828	0.434318	-0.114179	-0.016743	-0.177521	-0.106483
OnlineSecurity	0.389649	0.184743	0.300472	-0.025879	0.180349	0.141731
TechSupport	0.410403	0.158294	0.312395	-0.016865	0.182892	0.199585
PaperlessBilling	0.175638	-0.290589	0.101442	-0.000486	0.019840	0.020805
MonthlyCharges	0.684026	-0.583681	0.222660	-0.013865	0.191741	0.153185
TotalCharges	0.964147	0.029664	0.072782	-0.027704	-0.001717	0.006138
DSL	-0.109832	0.536456	0.657241	-0.101238	0.231369	0.167740
Fiber optic	0.414751	-0.793525	-0.172990	0.043232	-0.005967	-0.014795
oneY_contr	0.159970	0.098825	0.163573	-0.054545	0.077909	-0.470745
twoY_contr	0.372806	0.538287	-0.380227	0.089448	-0.177852	0.633042
CreCard	0.200756	0.197339	0.094900	0.945963	0.073438	-0.080998
Banktran	0.192544	0.130370	-0.053912	-0.346280	-0.065282	0.023709
tenure over 24	0.757305	0.387145	-0.075874	-0.028403	-0.188149	-0.239876
TO24INT	0.842096	0.442478	-0.116015	-0.021837	-0.212314	-0.134334
month over 84	0.655770	-0.593706	-0.085105	0.027590	0.108487	0.082926
MON84INT	0.681156	-0.592045	-0.085981	0.026219	0.107326	0.084338
total over 1500	0.836390	0.067833	0.170297	-0.058165	-0.016336	-0.094308
TOCH1500	0.952512	0.013729	0.107878	-0.037698	0.001794	0.001936

Reference

BlastChar. (2018, February 23). Telco Customer Churn. Retrieved from https://www.kaggle.com/blastchar/telco-customer-churn