Telecom Customer Churn Prediction

Springboard Data Science Capstone Project

Ashley Jiangyang

April 2020

Table of Contents

1 Introduction

The telecom industry continues to confront growing pricing pressure worldwide. While regional differences apply, wireless expansion is reaching a saturation point across multiple markers. In addition, the longstanding ability to diversify products and services based on handset selection and network quality is wearing off, and product lifecycles are shortening. Simultaneously, wireline businesses are facing competition from cable operators and a risking of disruption from the OTT (Over-the-top media service) players. All of these powerful trends are forcing telecom companies to respond through more competitive offers, bundles, and price cuts.

Given these challenging industry dynamics, managing customer base to reduces churn should be among any senior telecom executive's highest priorities. In this project, we will use the telecom churn dataset to exploring the product and services dynamic develop a comprehensive view of the customer and link that view to strategy-making recommendations. Furthermore, we will build machine learning models to predict customer behavior, which can be integrated into an agile test-and-learn process to the company to customize offers for different individual microsegments, which will be helpful to improve the customer retention.

2 Data Acquisition and Cleaning

The dataset is acquired from Kaggle: Telco Customer Churn – Focused Customer Retention Programs, which contains information about,

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

We drop the missing value (0.16%) and the Customer ID column. More details on the data cleaning process can be found in this IPython Notebook. The cleaned data is then ready for exploration. A detailed codebook for the data is given below.

CODE BOOK

Feature	Type	Counts	Missing Value	Statistical Summary	Info Area	Description
Gender	Category	7,032	0%	Male: 54% Female: 50%	Customer Demographic	Whether the customer is a male or a female
Senior Citizen	Category	7,032	0%	Yes: 16% No: 84%	Customer Demographic	Whether the customer is a senior citizen or not
Partner	Category	7,032	0%	Yes: 48% No: 52%	Customer Demographic	Whether the customer has a partner or not
Dependents	Category	7,032	0%	Yes: 30% No: 70%	Customer Demographic	Whether the customer has dependents or not
Partner	Category	7,032	0%	Yes: 48% No: 52%	Customer Demographic	Whether the customer has a partner or not
Tenure	Numeric	7,032	0%	Yes: 30% No: 70%	Customer Account Information	Number of months the customer has stayed with the company
Contract	Category	7,032	0%	Month-to-month: 55% One year: 21% Two year: 24%	Customer Account Information	The contract term of the customer
Paperless Billing	Category	7,032	0%	Yes: 59% No: 41%	Customer Account Information	Whether the customer has paperless billing or not
Payment Method	Category	7,032	0%	Electronic check: 33% Mailed check: 23% Bank transfer (automatic): 22% Credit card (automatic): 21%	Customer Account Information	The customer's payment method
Monthly Charges	Numeric	7,032	0%	Min: 18 25%: 36 50%: 70 75%: 90 Max: 119 Mean: 65	Customer account information	The amount charged to the customer monthly
Total Charges	Numeric	7,032	0%	Min: 19 25%: 401 50%: 1397 75%: 3994 Max: 8648 Mean: 2283	Customer account information	The total amount charged to the customer
Phone Service	Category	7,032	0%	Yes: 90% No: 10%	Customer Services	Whether the customer has a phone service or not
Multiple Lines	Category	7,032	0%	Yes: 42% No: 48% No phone service: 1%	Customer Services	Whether the customer has multiple lines or not
Internet Service	Category	7,032	0%	Fiber optic: 44% DSL: 34% No: 22%	Customer Services	Customer's internet service provider
Online Security	Category	7,032	0%	Yes: 28% No: 50% No internet service: 22%	Customer Services	Whether the customer has online security or not
Online Backup	Category	7,032	0%	Yes: 34% No: 44% No internet service: 22%	Customer Services	Whether the customer has online backup or not
Device Protection	Category	7,032	0%	Yes: 34% No: 44% No internet service: 22%	Customer Services	Whether the customer has device protection or not
Tech Support	Category	7,032	0%	Yes: 29% No: 49% No internet service: 22%	Customer Services	Whether the customer has tech support or not
Streaming TV	Category	7,032	0%	Yes: 38% No: 40% No internet service: 22%	Customer Services	Whether the customer has streaming TV or not
Streaming Movies	Category	7,032	0%	Yes: 39% No: 39% No internet service: 22%	Customer Services	Whether the customer has streaming movies or not

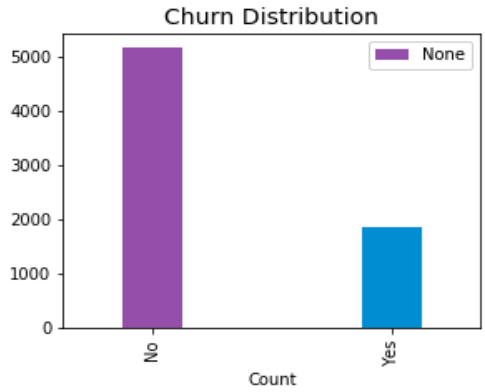
3 Data Exploration

3.1 Introduction to the Cleaned Data

There are 7,032 records and 20 features in the cleaned dataset. In this project, we use all the info from the dataset. We will go through each perspective of the information - Customer Demographics, Customer Account Information, and Customer Services – respectively with the churn dynamics among all the features in each domain. And a multi-features examination is also included in the analysis. Details about the Exploratory Data Analysis can be found in this IPython Notebook.

3.2 Customer Churn

Customer churn, also known as customer attraction or customer turnover, occurs when customers stop doing business with a company or stop using a company's services. By monitoring the churn dynamics, the companies can take advantage of that data to determine their customer retention success and identify strategies for improvement.



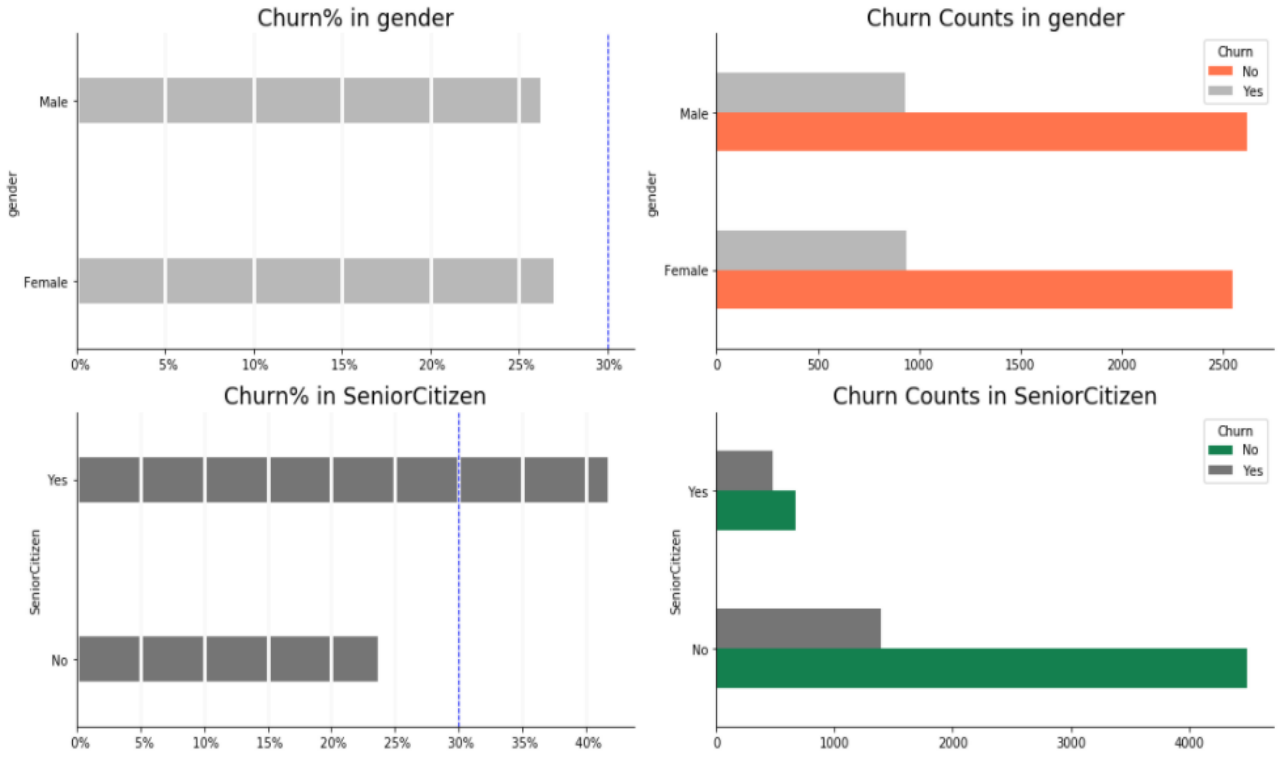

```
------- Churn Statistics -------
Churn Counts:
No     5163
Yes    1869
Name: Churn, dtype: int64

Churn Percentage:
No     0.734215
Yes    0.265785
Name: Churn, dtype: float64
```

In our data of telecom company, since we have no access of the exact time point that a customer joins the company and terminates the services, we would not be able to access the churn rate

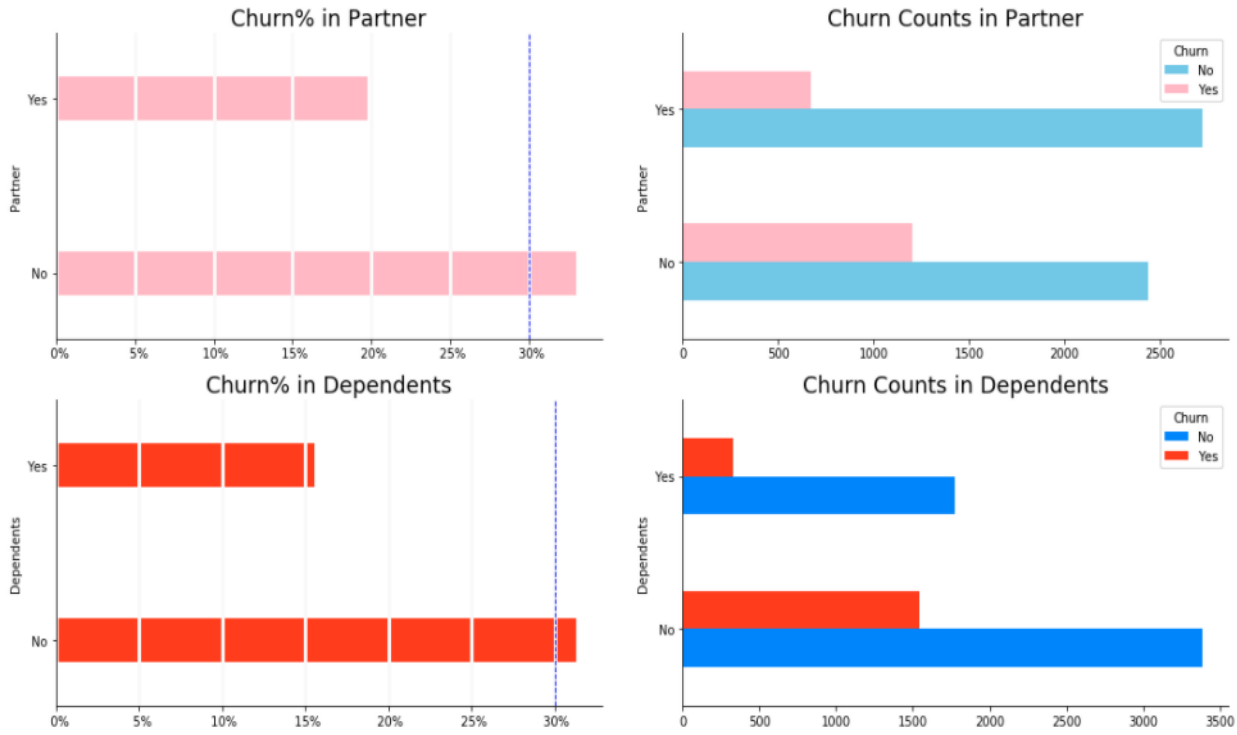
here. Instead, we calculate the churn percentage of the total population, which is about 26.6%. Assuming all the customer joined the business and the same period, and there's 73.4% of the customer is still staying with the business at the end of the period.

3.3 Customer Demographics

There are 4 features in this section – Gender, Senior Citizen, Partner and Dependents. We plot the percentage of churn and the count of churn breakdown by the category of each feature.



For both males and females, the churn rate is about ~26% and the customer of female and male is about the same counts. And for senior citizens, the churn percentage is over 40% whereas the churn percentage is much lower (~ 24%) for those who are not. A closer look at the count number reveals that there are substantial customers with the company is younger people, those who under the age of 62.



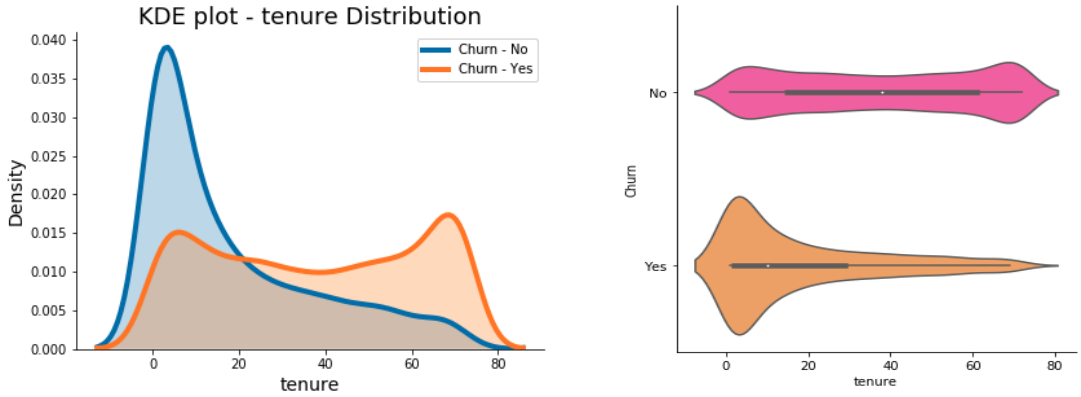
For those who without a partner, the churn percentage is higher (35% vs. 19%), which suggests that the partner plan might offer more cost-effective services. And for the dependents, those who have dependents have nearly half the churn percentage (16% vs. 33%) of those who don't.

```
Chi-Square Test Results

------- gender ------
P-value = 0.4737

------- SeniorCitizen ------
P-value = 0.0

------- Partner ------
P-value = 0.0

------- Dependents ------
P-value = 0.0
```

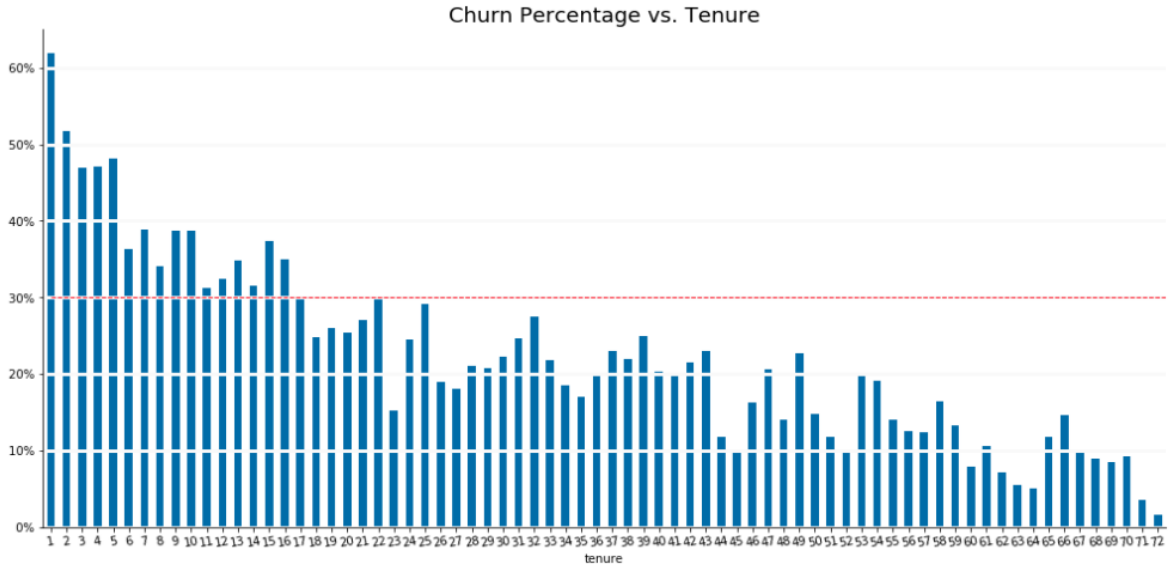
After visualization, we apply the Chi-square Test on each feature. The results suggest that there is a statistically significant difference in churn percentage when each feature from the Senior Citizen, Partner and Dependents is taken into account respectively.

3.4 Customer Account Information

From the customer account information perspective, the relationship between Churn and Tenure, Monthly Charge and Total Charge will be explored. Also, the churn difference between three category business features contract, paper billing and payment method will be display later.

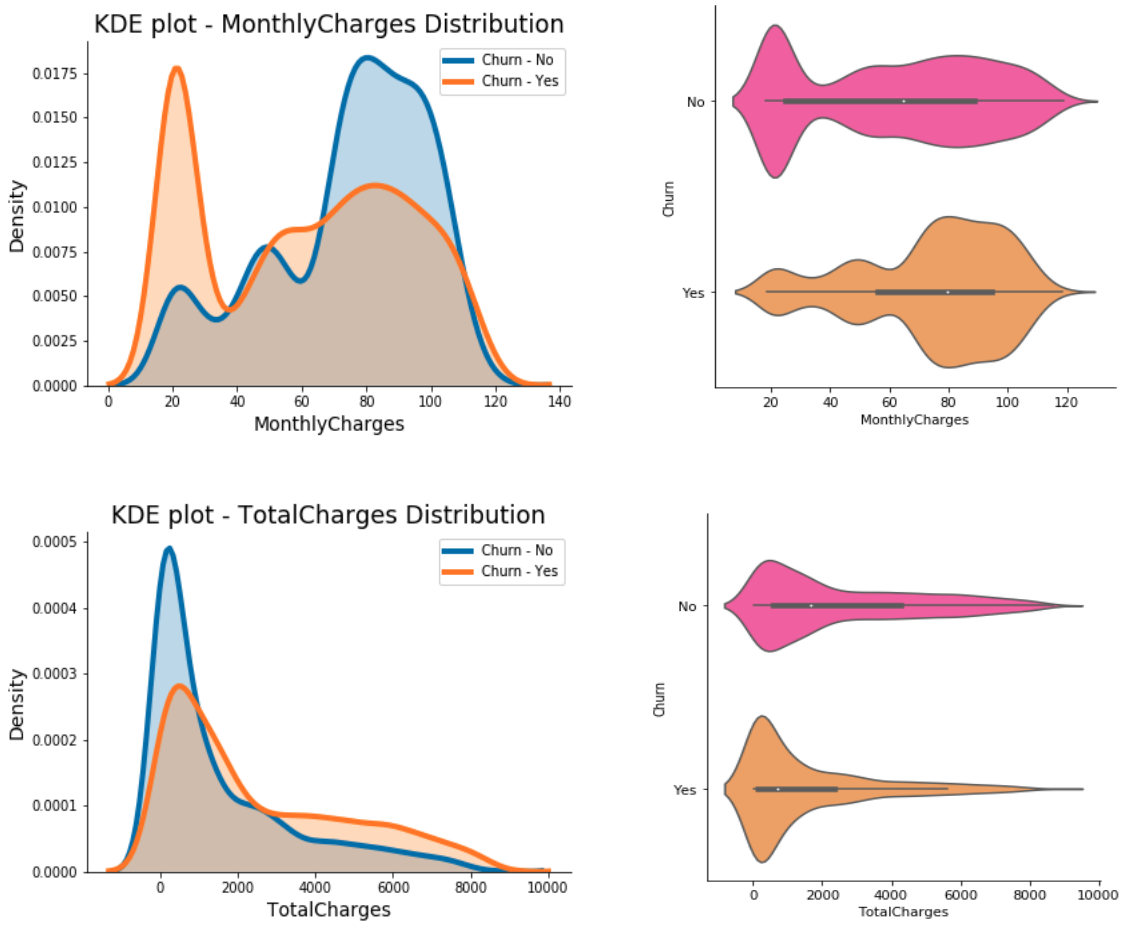


The non-churn group has reached its peak nearly the 7 months as their life span, which means that the company has a substantially newly joined customer – it's a positive signal that the business stays in health management that keep attracting customers. Then when we are shifting gear to look at the churn group, their tenures are more evenly distributed from 5 months to 70 months, that indicates people still stop their services even though they have been staying a long time with the company, maybe the company might need to consider how to maintain their customer by improving their services and products.



As complementary information, we plot the percentage of churn along with the tenure, within 17 months the churn percentage is as high as over 30%. And as the tenure increase, the churn percentage goes down. Look at the customer life span within 5 months - the churn percentage goes up high to 60%.

Let's take a look at the charge for both churn and non-churn group.



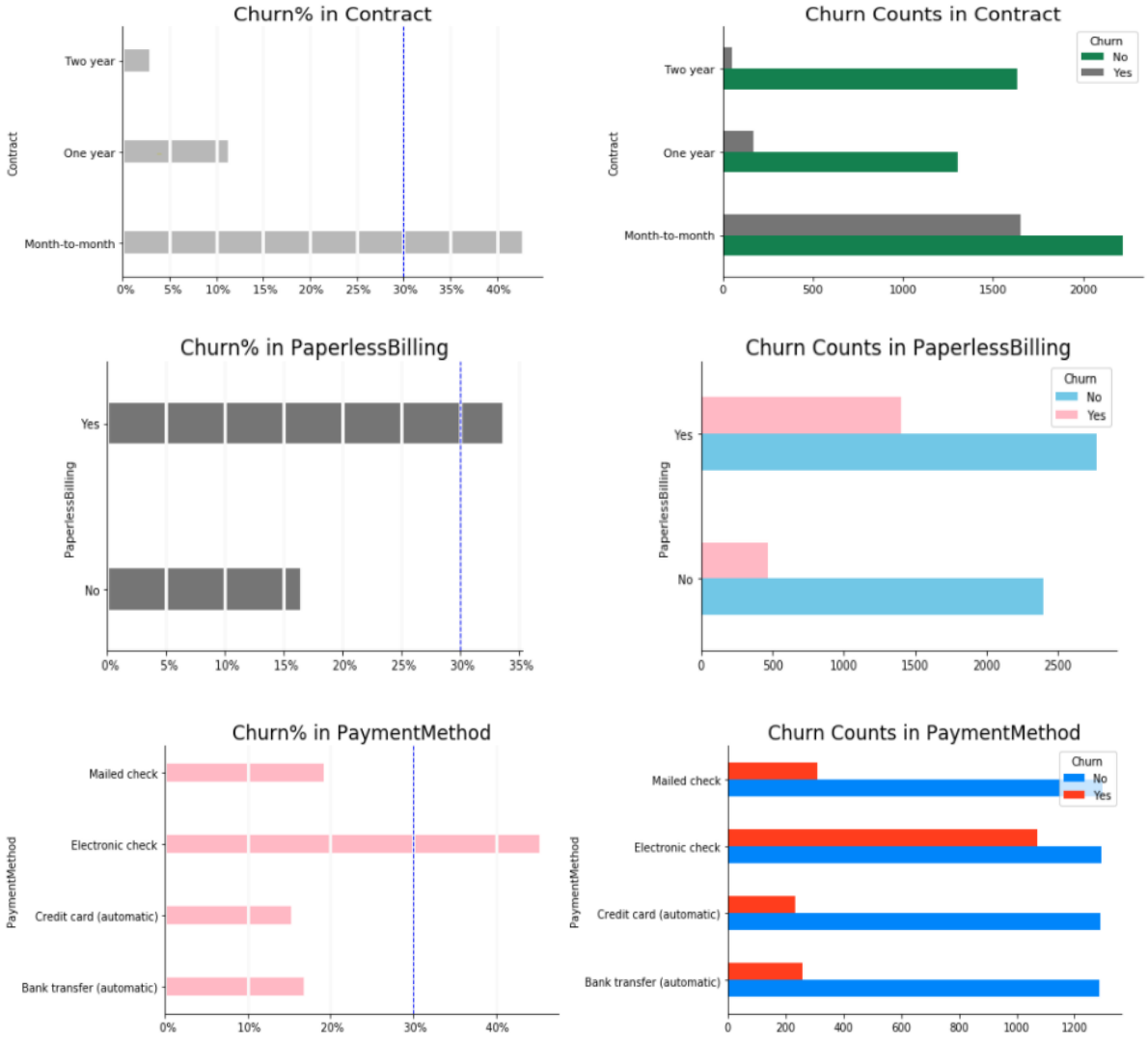
The monthly change reveals that the non-churn group has more people with higher charge and the total charge suggest the reverse pattern – during the peak around total charge as $1,000 there are fewer people in the non-churn group and when the total charger goes up above $2,000 towards the end (about $10,000), the non-churn group have more people gain more weights in this area.

Bootstrap hypothesis test for Churn vs. Non-Churn Group:

```
Hypothesis Test of tenure                | p-value = 0.0
Hypothesis Test of MonthlyCharges        | p-value = 1.0
Hypothesis Test of TotalCharges          | p-value = 0.0
```

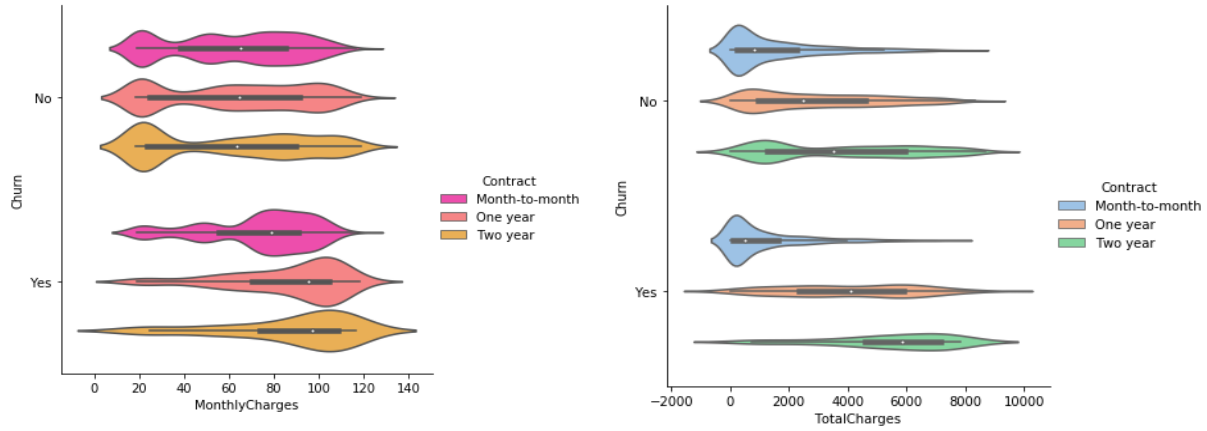
The bootstrap hypothesis testing suggests that there's a statistical difference of mean tenure in the non-churn and churn group, also the test results suggest statistical significance for monthly charge and total charge.

Let's then dive into more detailed customer account info.



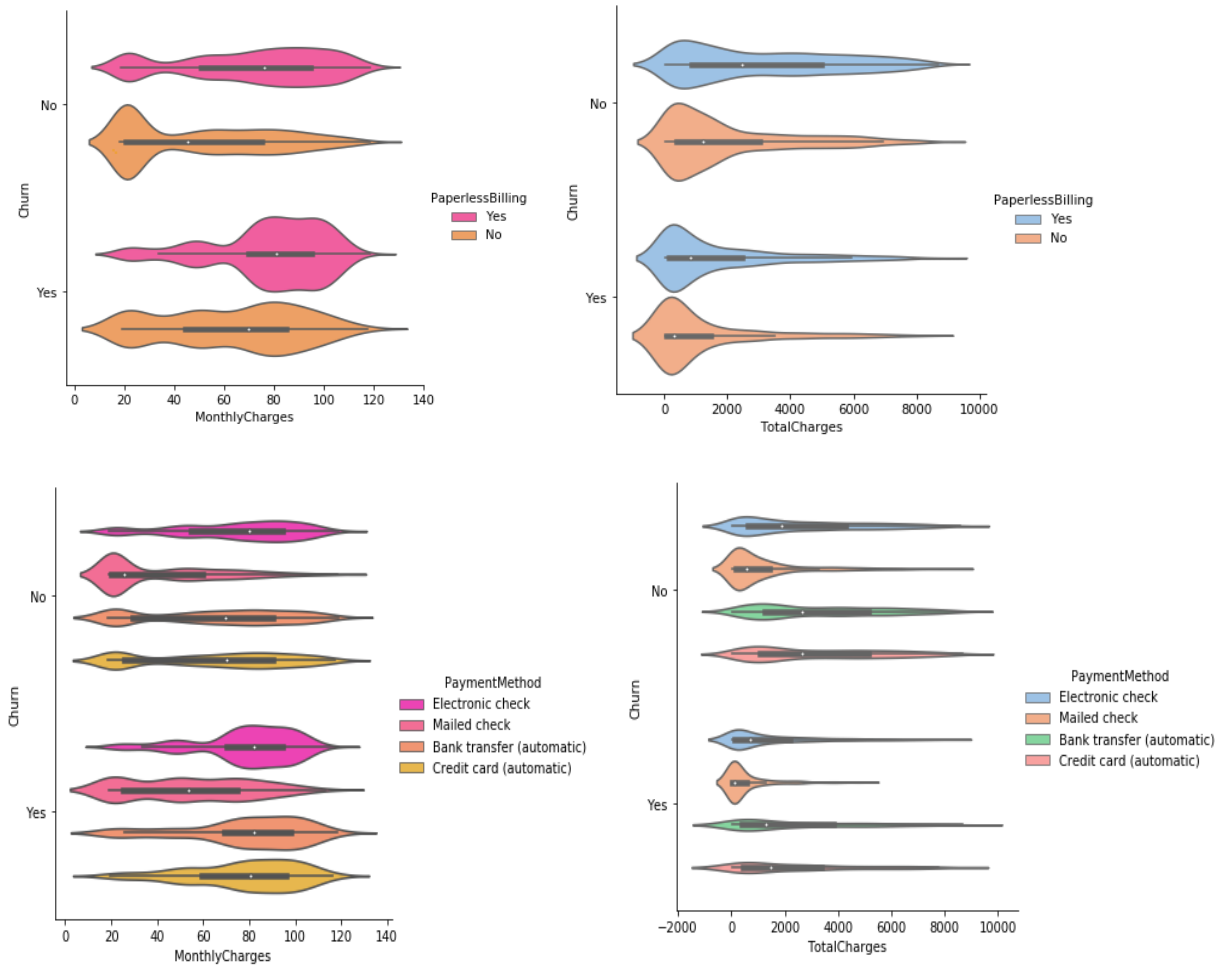
The longer the contract, the lower the churn percentage, noted that the month-to-month percentage is about 4 times higher (40%+) than the one year contract (12%), and he two-year contact has very good customer retention (churn percentage <5%), maybe the company offers more benefits/discount for their loyalty customers. The paperless billing group has about a 2-time high churn percentage (34% vs. 16%) of those not using paperless billing. The payment method shows the electronic check has much higher (43%) churn percentage – maybe it's hard to set up or cause inconvenience – the company should think about optimized this payment option. As complement info, the chi-square test indicates that the contract, paperless billing and payment method all are having a statistically significant impact on the churn.

For further information to understanding whether the charge effect those groups have a higher churn percentage, we t visualize the charge for those features.



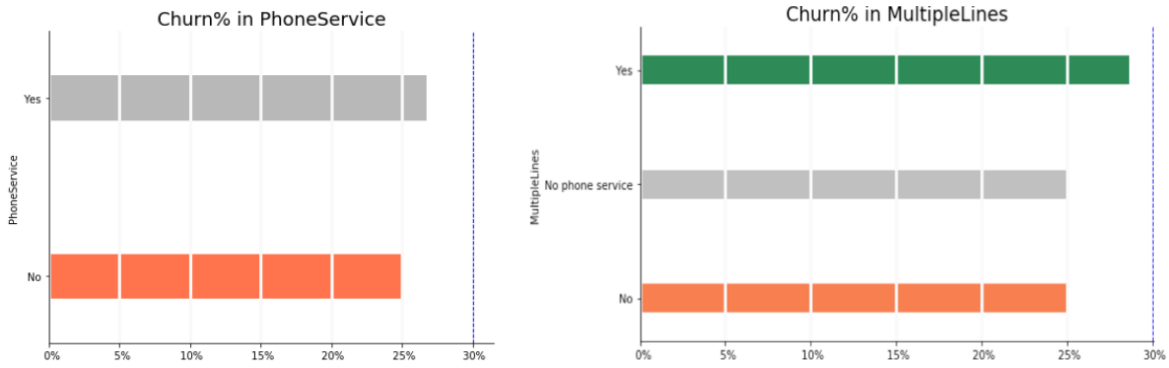
The monthly charge is higher for the churn group in general -which means it's highly likely that those people churn could be caused by unreasonable charge, most people have the monthly charge $60 - $120. And even for those in the non-churn group, we notice that there are more people have been charged in the $40 - $120 area.

This pattern also applies to the paperless billing and payment method – and very obvious in the monthly charge – if we take look at the monthly charge for people in the churn group, we notice that a great proportion of people fall into the range of relatively high monthly charge.

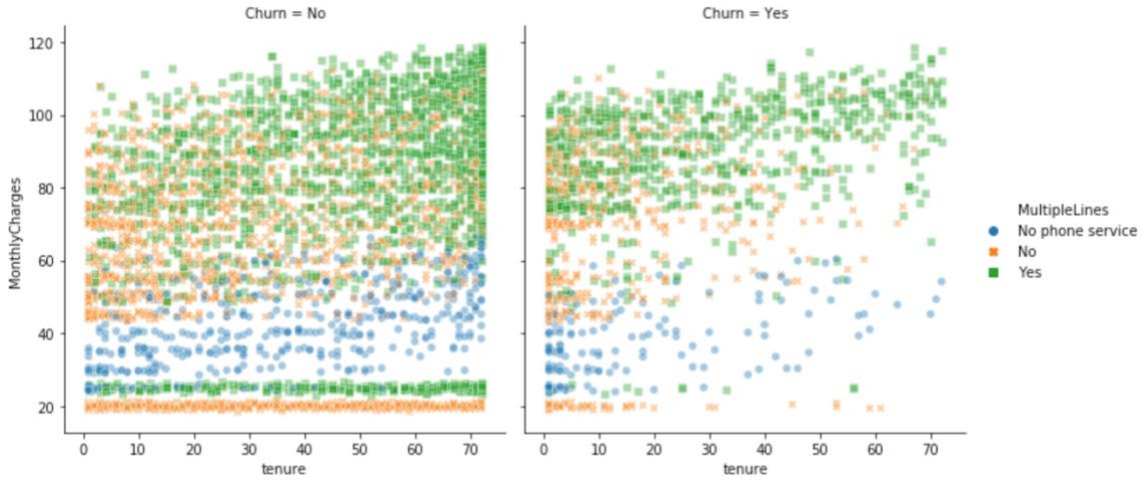


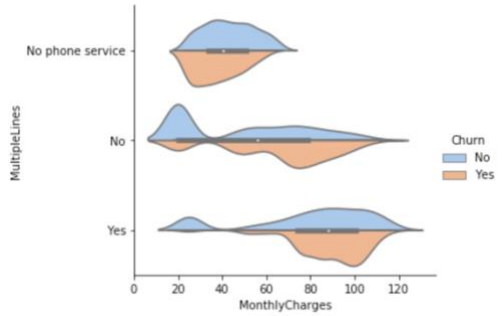
3.5 Customer Services

There are two main categories of customer services - **phone services** and **internet services**, for people who choose these services, they can have further options for more specialized services to opt-in. For people who stay with phone services, they can further choose multiline or not. And for those who have access to internet services, they are eligible to choose services like online security, online backup, device protection, tech support, streaming TV and streaming movies.



The churn percentage of people with phone services (27%) is a little bit higher than the people without (25%). And for those who have multiple lines, the churn percentage goes up to 28%. Which suggests the company man consider to improve the phone services and multiple services to prevent the churn percentage to increase.

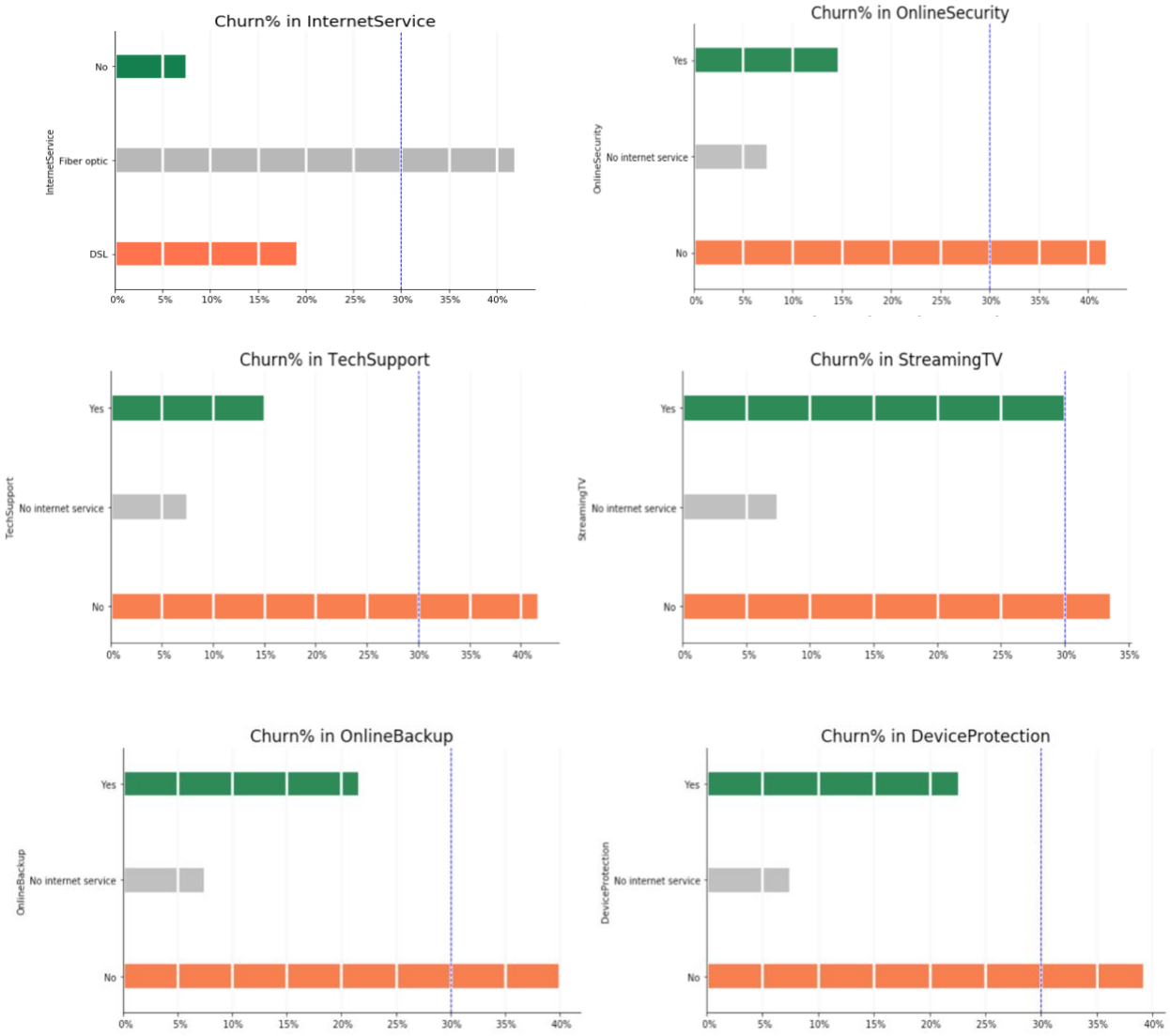


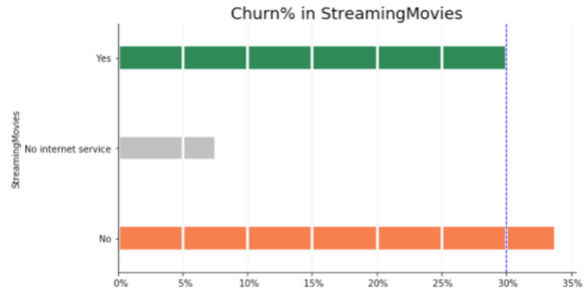


let's take a look at the monthly charge for multiline services. The charge for churn and the non-churn group seems not to have a huge difference. And either for phone services or multiple line services, the subscriber group does not have an obvious lower churn percentage, for the people who opt-in the multiple line services, the churn percentage is

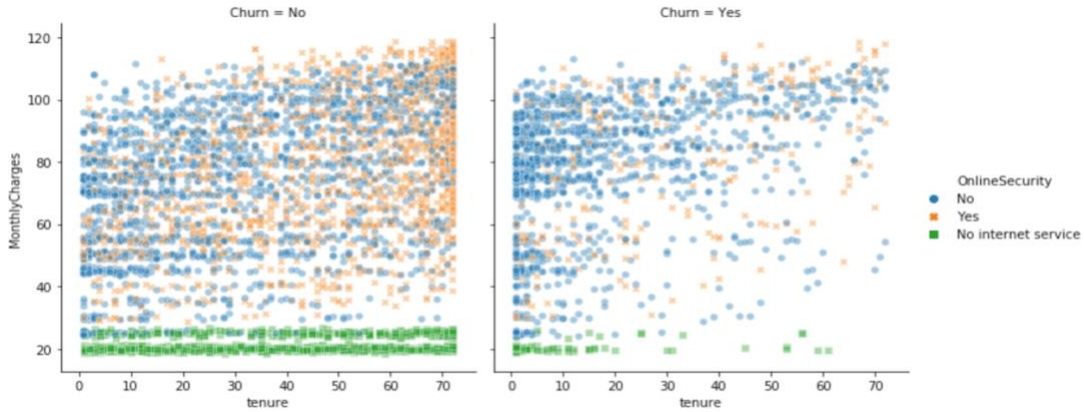
even higher. The monthly charge distribution is increasing along with the additive services people are choosing, which is reasonable. An interesting finding is that for the charge, for both churn or non-churn group, the multiple subscriber or non-subscriber tend to have more extreme charge whereas the charge for no phone services is centered in the middle-lower areas.

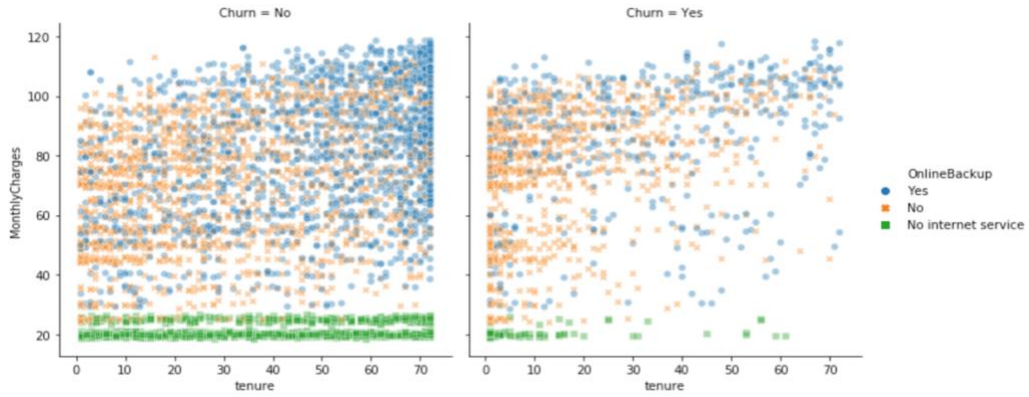
Let's take a look at the internet services.





For the internet services the fiber optic has the highest churn percentage (up to 42%) whereas the DSL is about half churn percentage (18%) of that, and the no internet services is the lowest (less than 8%). And for online security, online backup, device protection, tech support, streaming TV and streaming movies, the people who choose to have the services has lower churn percentage than who don't – hence, we would consider these company should recommend these services to their customer to encourage more engagement – especially for online security and tech support.





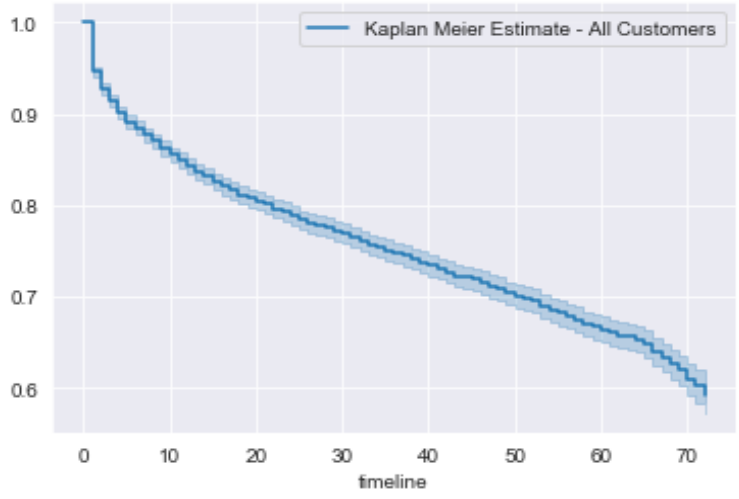
Additionally, we plot the monthly charge vs. tenure for online security and online backup, in the non-churn group, people with the services tend to cluster at longer - tenure area, and the charge is about in the same range whether the customer chooses online backup/security or not.
The chi-square test shows only the phone services is not statistically significantly associated with the churn.

4 Survival Analysis

In this section, we will incorporate a useful tool – **survival analysis** for customer churn analysis. Survival analysis can be used as an exploratory tool to compare the differences in customer lifetime between cohorts, customer segments, or customer archetypes. With the help of Survival Analysis, we can focus on churn prevention efforts of high-value customers with low survival time. This analysis also helps us to calculate Customer Life Time Value. In this use case, the Event is defined as the time at which the customer churns/unsubscribe. The time scale is represented by months. Given the data we have, we will conduct cohort analysis and a simple prediction model to help us understand customer retention.

4.1 The total customer cohort retention

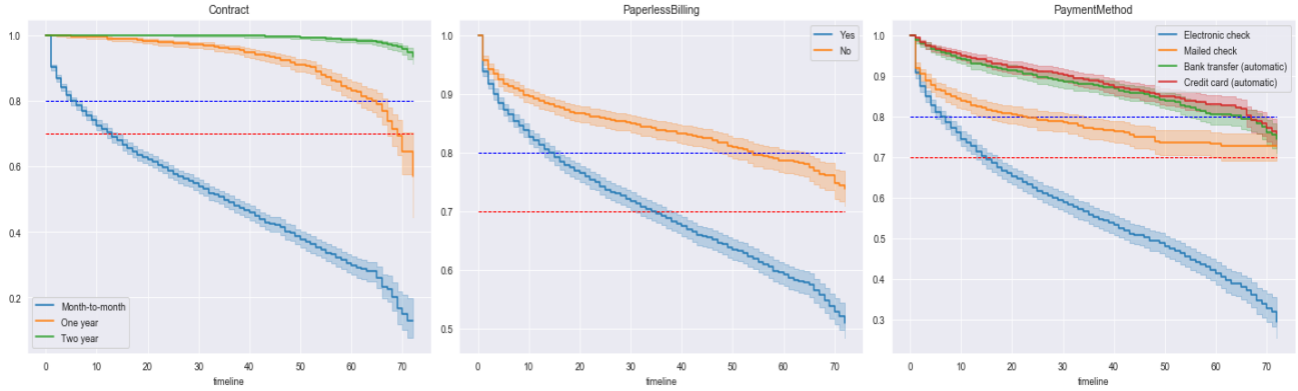
We apply the Kaplan – Meier estimate on the whole customer cohort. The y-axis represents the survival probability – the probability a customer is still staying with the telecom company after t month, where t month is on the x-axis (tenure).



Let's take a look at our plot when tenure is within 5 months, we see that there's a steep drop which indicates the churn rate is high, and along with the tenure increase, the churn rate is slowing down, and notice that the churn rate increase after 65 months. Overall, the company is doing a good job maintaining the customers – their probability a customer still with the company after 5 years (72 months) is about 0.59, and the probability a customer still stay with the company after 2 years is about 80%.

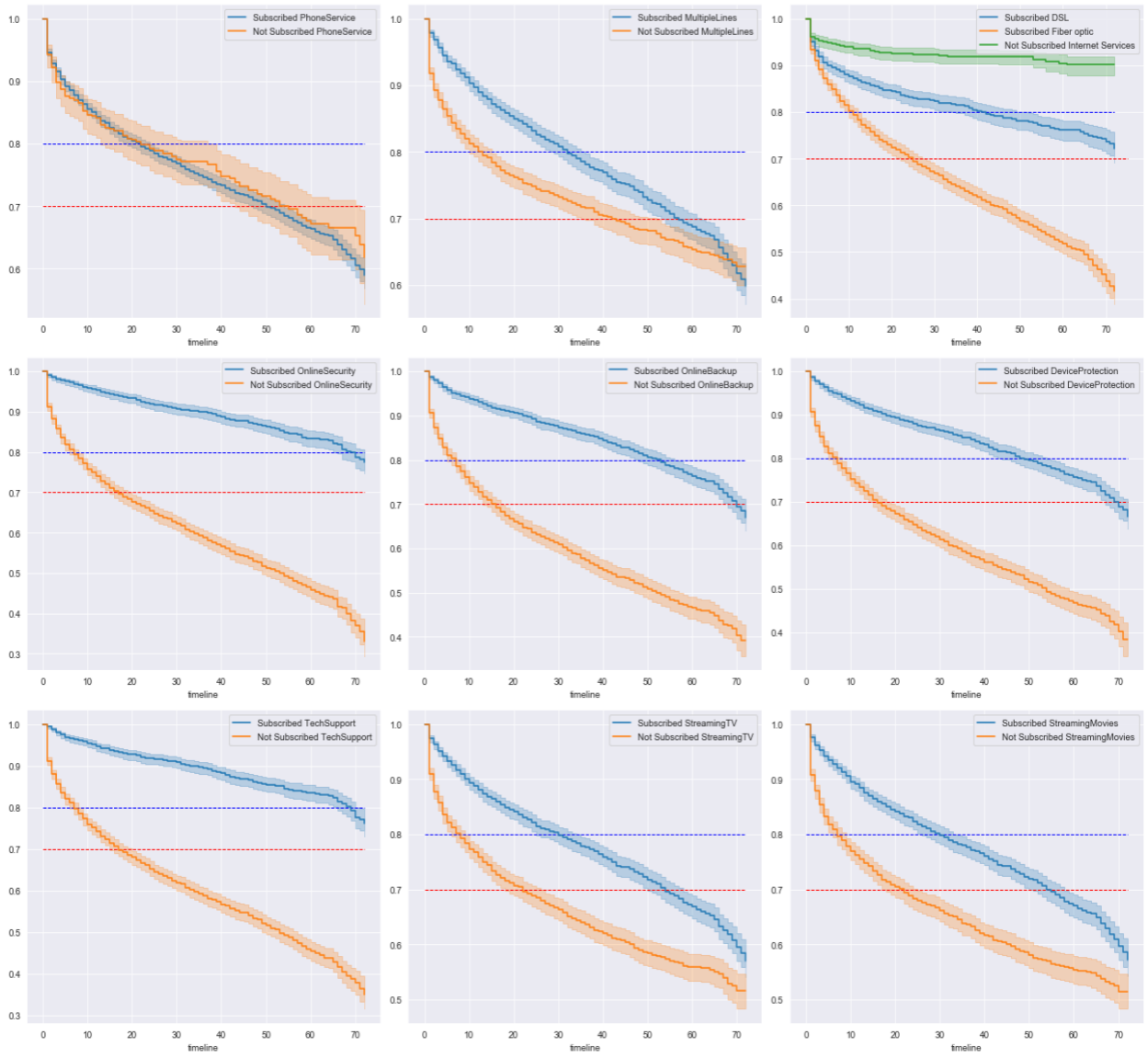
4.2 Cohort Analysis

In this section, we will look at the retention of the different cohorts that segmented by the business services types. More specifically, we are interested in each type of service - which business channel has better retention.



For contact, the one-year and two-year contract have better retention, and the customer with a two-year contract still has retention probability over 95% after 5 years, and the customer with a month-to-month contract only has 50% retention probability after 32 months. Customers who do not choose the paper billing have better retention. And for payment methods, the people who choose electronic check only have 30% retention probability whereas the three other method – mailed check, bank transfer and credit card remain about 75%, however, the trajectory of the mailed check is not the same from the other two – the retention probability decrease fast and it becomes more steady, whereas the other two have a more consistent decreasing curve.

We also include the survival curve for all the customer services, check the detailed plot below. The results are consistent with previous EDA session, the customer services all have remarkable work in retaining customers.



4.3 Prediction for individual customers by Cox Proportional Hazard Model

In this section, we build a Cox Proportional Hazard Model -we have each customer's tenure when they churned (the event time T) and the customer's Gender, Monthly Charges, Dependants, Partner, Phone Service, etc. The other variables are the covariates in this example. We are often interested in how these covariates impact the survival probability function. The Cox (proportional hazard) model is one of the most popular models combining the covariates and the survival function. It starts with modeling the hazard function.

$$h(t|X = x) = h_0(t)\exp(x^T\beta)$$

Here, β is the vector of coefficients of each covariate. The function ho(t) is called the baseline hazard function.

The Cox model assumes that the covariates have a linear multiplication effect on the hazard function and the effect stays the same across time.

From the above equation we can also derive cumulative conditional hazard function as below:

$$H(t|x) = \exp(x^T\beta)\int_0^t h_0(s)ds = \exp(x^T\beta)H_0(t)$$

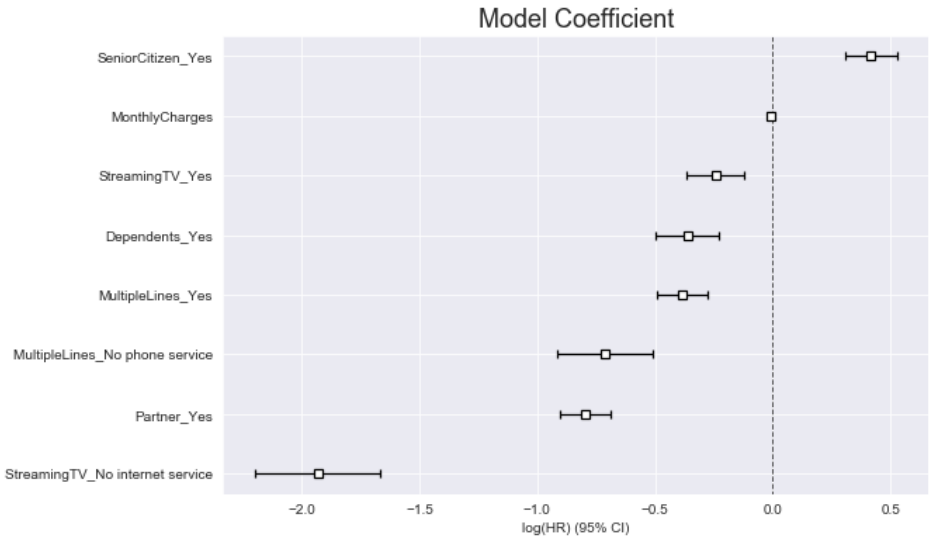
As we are already aware that we can derive survival function from the hazard function with the help of expression derived in the above section. Thus, we can get the survival function for each customer.

The model results:

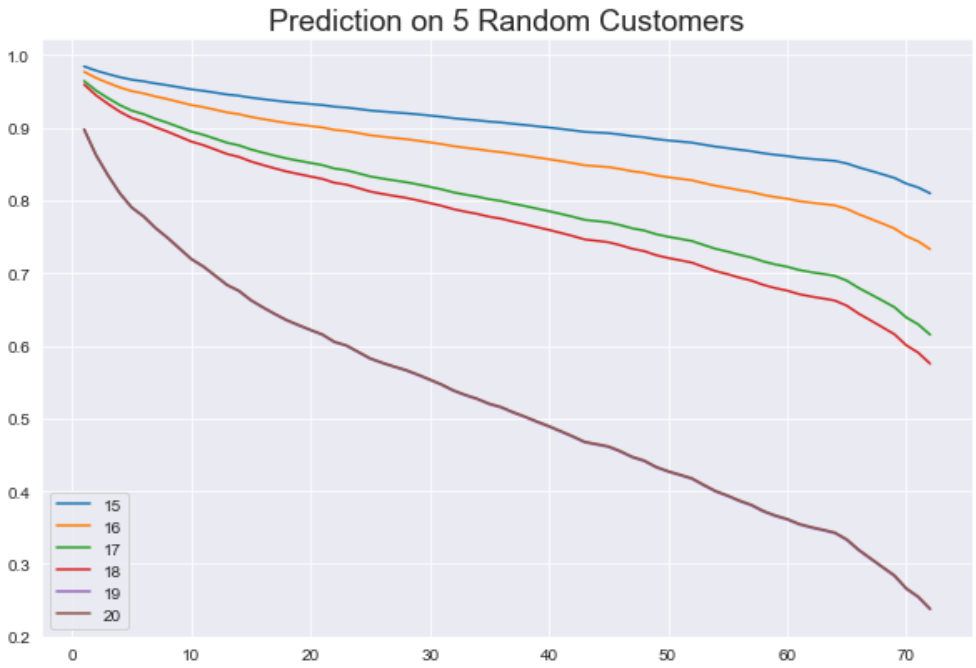
	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)
MonthlyCharges	-0.01	0.99	0.00	-0.01	-0.00	0.99	1.00	-3.00	<0.005	8.55
Partner_Yes	-0.80	0.45	0.05	-0.90	-0.69	0.41	0.50	-14.65	<0.005	158.96
Dependents_Yes	-0.36	0.70	0.07	-0.50	-0.23	0.61	0.80	-5.30	<0.005	23.04
MultipleLines_No phone service	-0.71	0.49	0.10	-0.91	-0.51	0.40	0.60	-6.86	<0.005	37.07
MultipleLines_Yes	-0.38	0.68	0.06	-0.49	-0.28	0.61	0.76	-6.94	<0.005	37.92
SeniorCitizen_Yes	0.42	1.52	0.06	0.31	0.53	1.36	1.69	7.53	<0.005	44.19
StreamingTV_No internet service	-1.93	0.14	0.14	-2.20	-1.67	0.11	0.19	-14.23	<0.005	150.26
StreamingTV_Yes	-0.24	0.78	0.06	-0.36	-0.12	0.69	0.89	-3.93	<0.005	13.54

The summary results show the coefficient and the P-values of the covariates in predicting the churn risk. All the covariates play a statistically significant role regarding the prediction model. Taking a look at these coefficients for a moment, Partner – Yes has a coefficient of about – 0.90, thus, If customers have partner, the risk of churn is exp(-0.90) = 0.45 times of those who don't – which mean the no partner group has a lower risk of churn.

Let's take a look at the coefficient in the plot.



The negative values indicate they have a lower risk of churn compared to the baseline group, in our model, only being senior citizen will have higher risk oh churn; Subscribing the streaming TV, multiline, not choosing internet services all have a lower risk of churn.



Finally, we randomly pick 5 individual from our dataset to predict their churn, Creating the survival curves at each customer level helps us in proactively creating a tailor-made strategy for high-value customers for different survival risk segments along the timeline.

5 Modeling

5.1 Data Pre-processing

Before fitting the data into any machine learning algorithms, there are a few pre-processing steps that should be performed on the data. We outline these steps below.

1. Dummy encoding: In the dataset, we will code the binary features as 0/1, and we will convert the categorical features into dummy features.

2. Data Splitting: The second step involves splitting the converted dataset into train and test dataset. In our project, we split the data into 80% - 20% ratio.

3. Resampling or Weighting: In the third step, we take care of the imbalanced data issue by addressing it at the data level either resampling the majority class examples or oversampling the minority class examples. There are three techniques will be used on the project, 1). Random Oversampling, 2). Synthetic Minority Oversampling Technique (SMOTE), 3).Weighting – More weights are given to minority class examples. We will explore these techniques within each type of classification algorithm and pick the best ones.

4. Scaling: We applying the MinMaxScaler from Sciki-Learn to scale all the features in the ranges of value [0, 1].

5.2 Modeling Pipeline and Evaluation Metric

After the preprocessing steps described above, we fit the training data into classification algorithms to build the models and test them on the testing data. We use pipeline function from Sciki-learn to combine the resampling/weighting, scaling, and initial classifier together, follow the pre-processing steps on the 80% data as the training set. Once we know the optimal hyperparameters and the best resampling/weighting techniques, we will refit the model on the whole dataset. For hyperparameter tuning, 5-fold cross-validation with grid search is applied. Given that we have the unbalanced data for target features, we won't be stick to much on accuracy. We want to have a high true positive rate (recall) with the churn tagged as positive classes while we also trying to lower the false positive rate as well – which means we want to have higher precision as much as it can in the meantime.

We then take the business problem into our consideration for metric priority – we would like to correctly capture all the true positive cases of churn customers as much as we can, we might have tolerance for some false positive cases – the non-churn customer that'll be identified as churn group incorrectly – instead of having a lot of false-negative cases, which means we will lose track of those churn customers who might our high-valued customers. Given this, we will stick to recall as the top priority to compare the models.

5.3 Logistic Regression

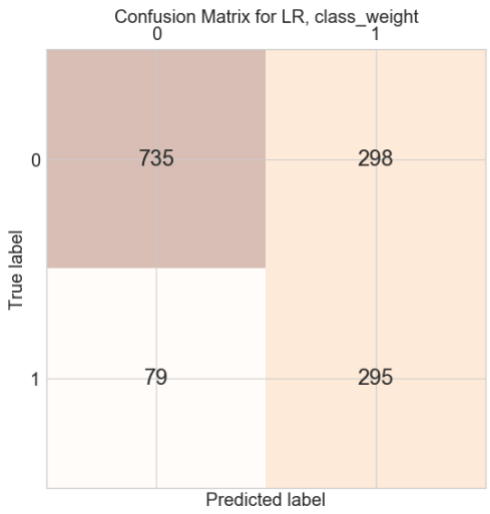
For logistic regression algorithm, we use all 4 pre-processing that describe in 5.1, 80% of the total data are used as the training dataset. Table 1 shows the results for various metrics for all different resampling/weighting techniques we used.

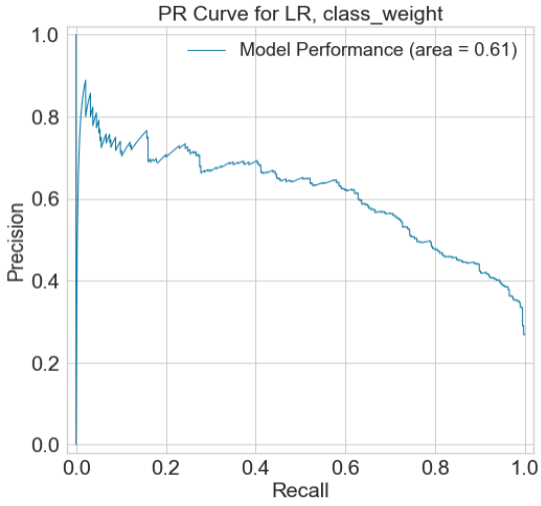
Table 1: logistic Regression Classifier – Choosing the best Resampling/weighting technique

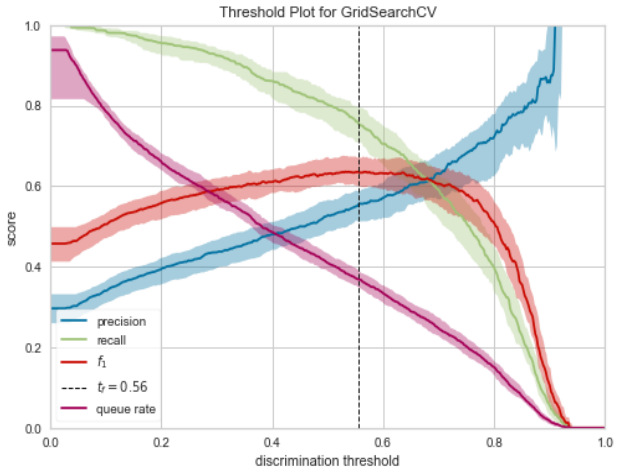
Resampling/weighting techniques	**Class**	**Metrics**			**ROC AUC**	**PR AUC**
		Precision	**Recall**	**F1-Score**		
RandomUnderSample	0	0.90	0.71	0.80	0.83	0.61
	1	0.50	0.78	0.61		
Weighting	0	0.90	0.71	0.80	0.83	0.61
	1	0.50	0.79	0.61		
SMOTE	0	0.90	0.71	0.79	0.83	0.61
	1	0.49	0.77	0.60		

In general, logistic regression has a generic performance, and the three sampling methods have almost the same performance. Look at the recall over the three resampling methods- we choose weighting here.

Let's combine the statistics with the plot to further look into our classifier performance. The PR AUC is 0.61, which is a mediocre performance. There is substantial misclassification non-churn group as the false positive, noted that for churn group – class 1, the precision is low and the recall is okay, which means that many non-churn customers will be classified as churn, and the recall – about 0.78, we will have some false negative – cases that's churn hasn't been – but we won't missing that much churn customers. In general, I would say the performance of logistics regression align with our business goal.







Noted that our use of binary of classification algorithms is to use the probability they produce to determine cases are churn or not churn – in default, the threshold of probability is 0.5 and any probability above that would be identified as churn (the positive case), otherwise, the negative class is. The discrimination threshold plot at the right corner is a visualization of precision, recall, F1 score, and queue rate with respect to the discrimination threshold of the binary classifier. The virilization tuned for an optimal threshold based on the F1 Score.

Recall our business goal - we are stick to the recall still, we might be thinking about lower the threshold a little bit to obtain a relatively higher recall at the expense of reduction of the precision, that mean we give the first priority to capture all the churn customers and we might include some unnecessary cases though.

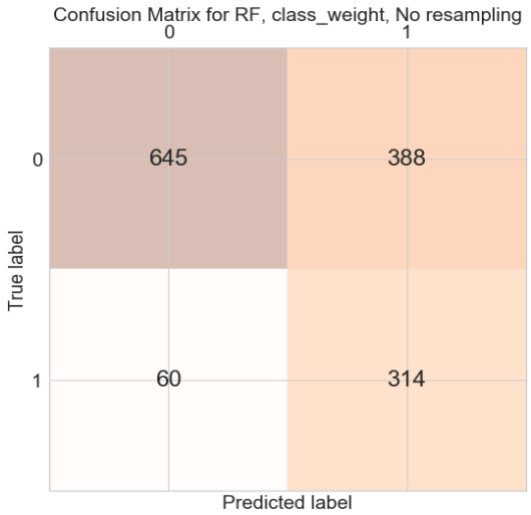
5.4 Gaussian Naïve Bayes

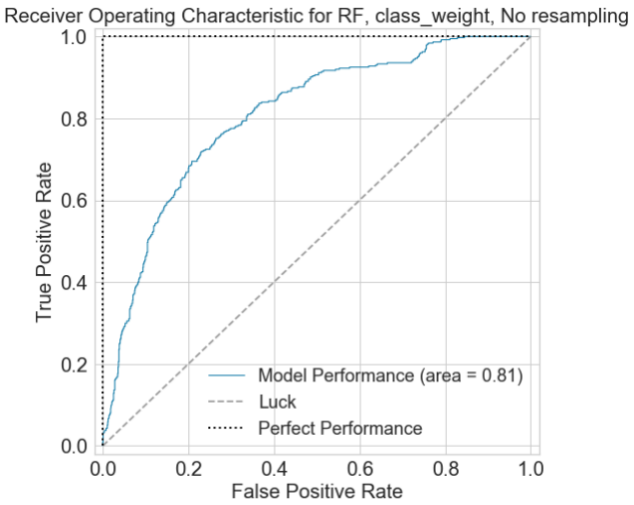
For Gaussian Naïve Classifier algorithm, we use all 4 pre-processing that describe in 5.1, 80% of the total data are used as the training dataset. Table 2 shows the results of various metrics for all different resampling/weighting techniques we used.

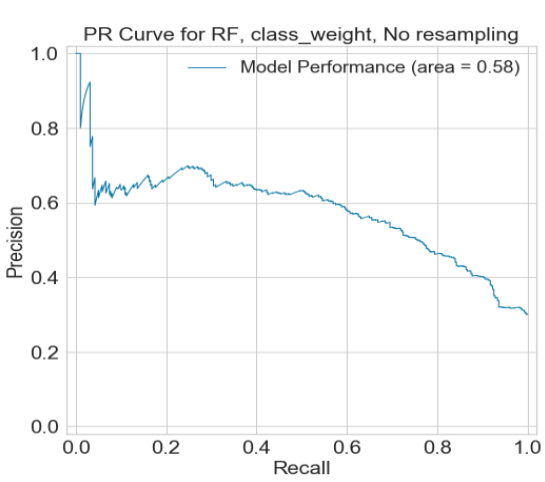
Table 2: Gaussian Naïve Bayes Classifier – Choosing the best Resampling/weighting technique

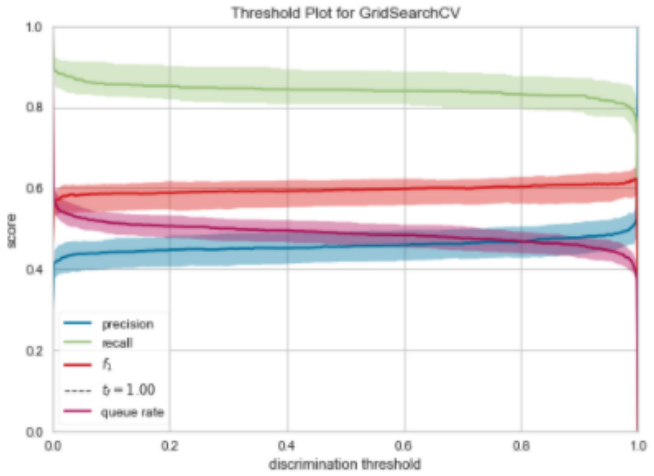
Resampling/weighting techniques	Class	Metrics			ROC AUC	PR AUC
		Precision	Recall	F1-Score		
RandomUnderSample	0	0.91	0.60	0.73	0.81	0.58
	1	0.43	0.84	0.57		
Weighting	0	0.91	0.62	0.74	0.81	0.58
	1	0.45	0.84	0.58		
SMOTE	0	0.91	0.64	0.75	0.81	0.58
	1	0.45	0.83	0.59		

The Gaussian Naïve Bayes is facing the similar problems, The ROC AUC and PR AUC is slightly worse than the logistic regression, the precision is low – which means we have lots of people are classified into churn whereas they are not. And the recall is improved. Compared to logistic regression, the gaussian naïve bayes model has a slightly better performance regarding aligning with our business goal. Which indicates we are having less false negative cases.









Then we also take a look at the discrimination threshold plot - all the metrics look consistent along with the change of the threshold. And we only see a slight change at the two extreme values, so we can use the default value here.

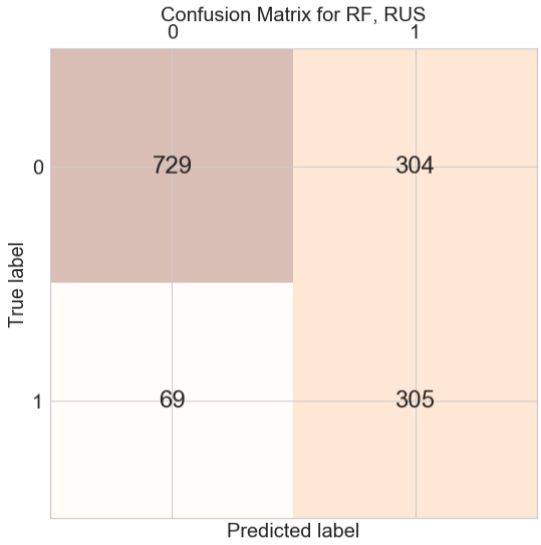
5.5 Random Forest

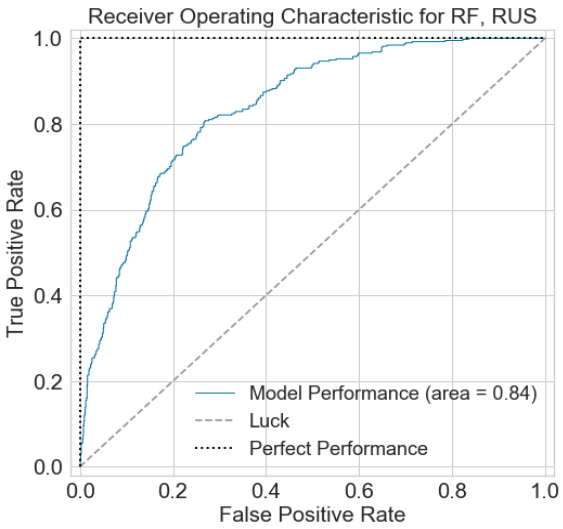
For Random Forest Classifier algorithm, we use all 4 pre-processing that describe in 5.1, 80% of the total data are used as the training dataset. Table 3 shows the results of various metrics for all different resampling/weighting techniques we used.

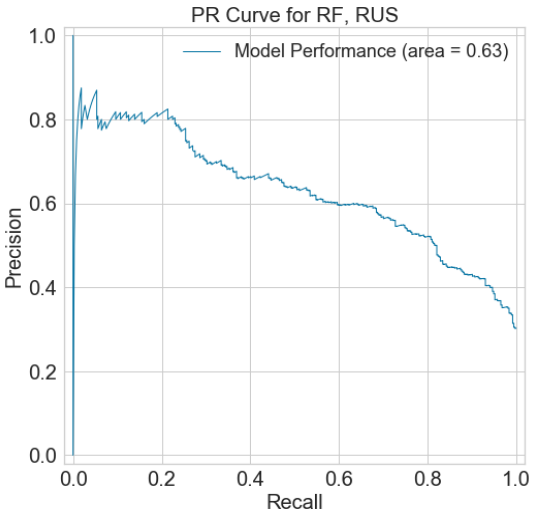
Table 3: Random Forest Classifier – Choosing the best Resampling/weighting technique

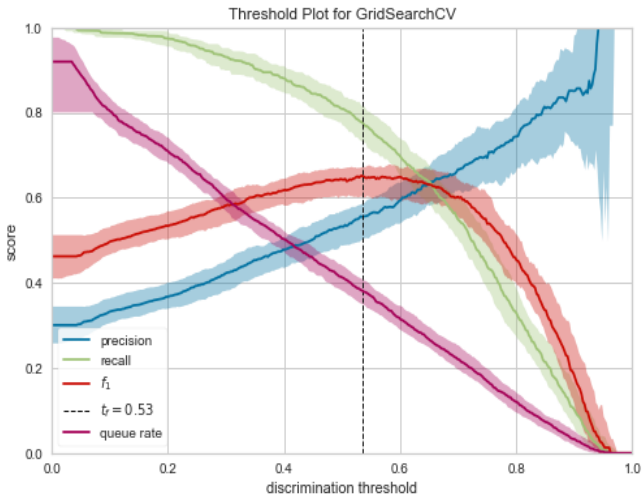
Resampling/weighting techniques	**Class**	**Metrics**			**ROC AUC**	**PR AUC**
		Precision	**Recall**	**F1-Score**		
RandomUnderSample	0	0.91	0.70	0.79	0.84	0.63
	1	0.50	0.82	0.62		
Weighting	0	0.91	0.72	0.80	0.84	0.64
	1	0.51	0.80	0.62		
SMOTE	0	0.86	0.86	0.84	0.83	0.62
	1	0.58	0.63	0.59		

The random Forest gives us a close performance to then naïve bayes above – the classifiers product a high false positive. And both ROC AUC and PR AUC have a slight improvement compared to naïve Bayes and logistic regression.









Note that the PR AUC is getting better here – though it's still not a great value.

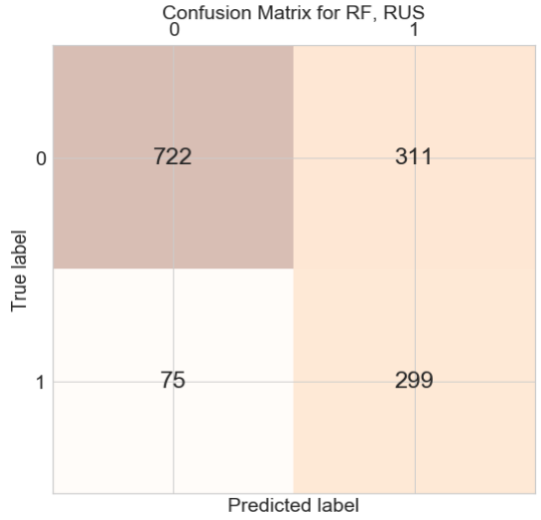
For discrimination threshold, we wouldn't want to change the threshold value – since we already have a decent recall, and we also notice that the precision score decrease fast along with the decrease of thresholds, we don't want to have a too-small precision score as well if we shift the threshold towards to the left side.

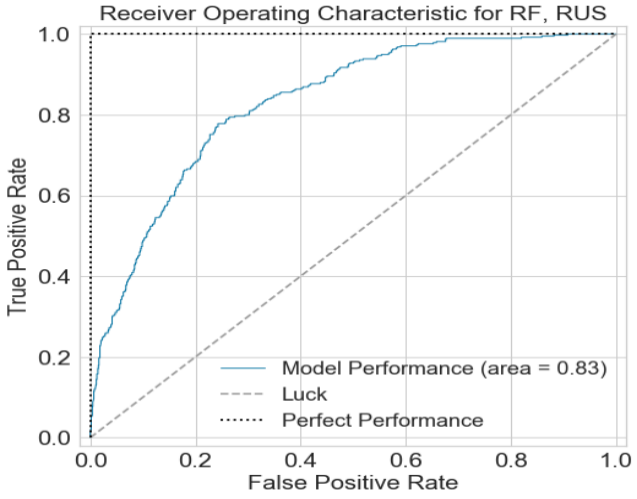
5.6 XGboost

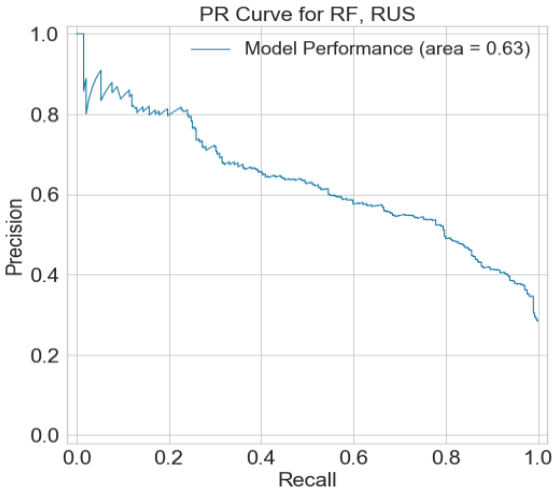
For Xgboost Classifier algorithm, we use all 4 pre-processing that describe in 5.1, 80% of the total data are used as the training dataset. Table 3 shows the results of various metrics for all different resampling/weighting techniques we used.

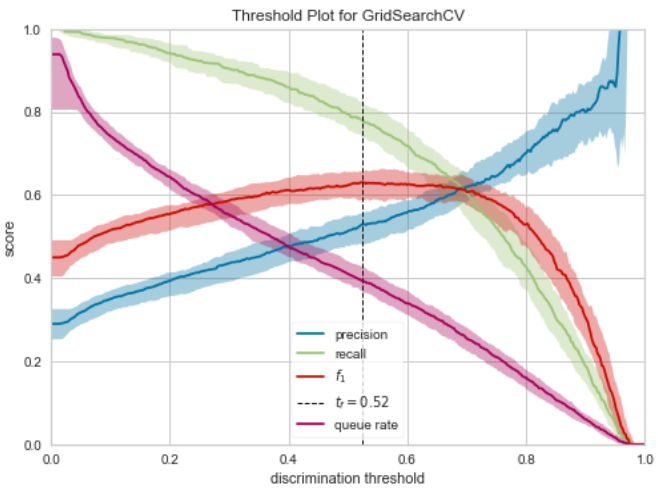
Table 4: XGBoost Classifier – Choosing the best Resampling/weighting technique

Resampling/weighting techniques	Class	Metrics			ROC AUC	PR AUC
		Precision	Recall	F1-Score		
RandomUnderSample	0	0.91	0.70	0.79	0.83	0.63
	1	0.49	0.80	0.61		
Weighting	0	0.84	0.89	0.86	0.84	0.66
	1	0.63	0.55	0.59		
SMOTE	0	0.85	0.87	0.86	0.83	0.64
	1	0.62	0.58	0.60		









Noted that the results for XGboost are very close to random Forest – it has a precision of 0.49 and recall as 0.80. And PR AUC and ROC AUC have the exact same value as random forest. The optimized threshold is about 0.52 – which aligns with our business goal - we obtain a relatively high recall and also taking the precision into account.

5.7 Model Comparison

To compare across all the different algorithms, we plot the PR AUC and ROC AUC for all the models and summarized the metrics into the table below.

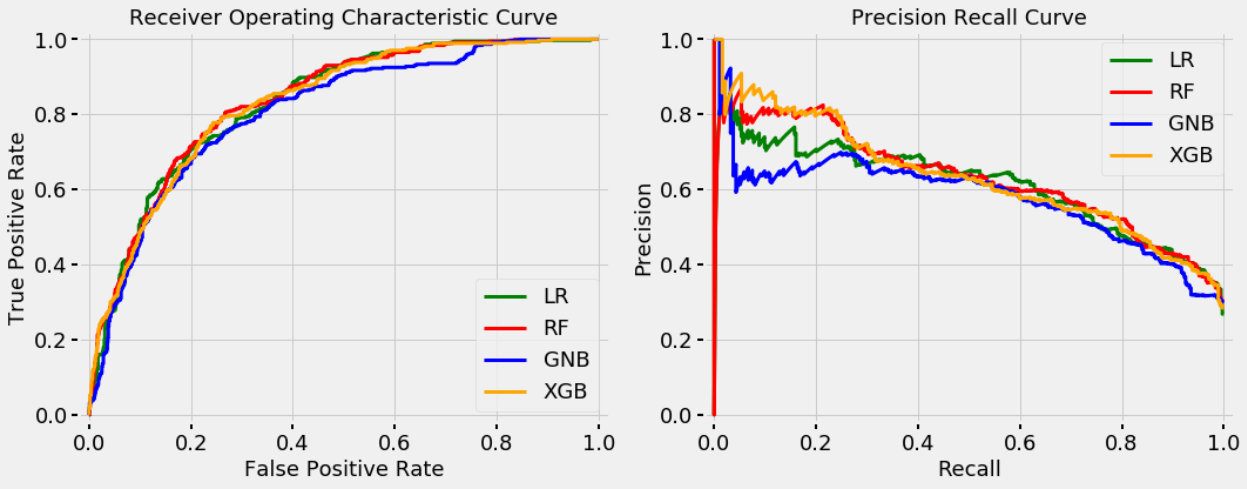


Table 5: Model Comparison

Algorithm	Sampling Method	Class	Metrics			ROC AUC	PR AUC
			Precision	Recall	F1-Score		
Logistic Regression	weighting	1	0.50	0.79	0.61	0.83	0.61
Gaussian NB	weighting	1	0.45	**0.84**	0.58	0.81	0.58
Random Forest	Random UnderSample	1	0.50	0.82	0.62	0.84	0.63
XGBoost	Random UnderSample	1	0.49	0.80	0.61	0.83	0.63

For each algorithm, we mainly compare on recall and the Gaussian naïve Bayes gains the highest value of recall - we would choose it as our final model.

This model will have the best performance in terms of capturing the true positive out of the truly positive cases – we definitely want to keep track of the high-value customer who is highly likely to churn as much as we can. However, we have a relatively poor precision at the expanse, which means when we apply our model to the current customer for prediction, we will have a substantial prediction of churn that they won't churn actually.

6 Limitation and Future Work

In summary, with all the algorithms we use above, the performances of the resampling method vary when it comes to combining different algorithms. The class weight works better with logistic regression and Gaussian Naïve Bayes, and the random oversampling works better when we use the random Forest and XGboost. Given that our unbalanced data, all the models are relatively good to identify the negative class – non-churn group, when it comes to the churn class, the performance is not so good. In a business sense – improving the separation ability of the model will be the priority for the further work, we may consider,

1) Acquire more data to train the classifier, we only have about 7,032 observations to fitting the algorithm. Gaining more data can help us capture more information.

2) Acquire more feature information about the customer and the subscription channel, such as the date that a customer joins and churn. Besides, information such as customer engagement/ activities would be more insightful – how long/frequent does a customer using steam TV, etc.

3) Conduct user experience research, to gaining more information about the reason why the customer likes or dislikes the specific type of services, which can help us get more information on the quality of the services. We can use that insight to conduct the campaign to keep and re-attract high-value customers.