**CUSTOMER CHURN ANALYSIS**

**Background**

* Telecom companies face major challenge with customer churn, as customers switch to alternate provider due to various reasons like lower cost, multi (combo) service offerings, marketing promotions by competitors, etc.
* In the USA, there are 7 telecom companies that serve customers in all 50 states. 13 regional companies serve 29 plus states and thousands of local companies provide internet and TV broadcasting services**#**.
* Identifying these potential customers early on who may voluntarily churn and providing them retention incentives in form of discounts & combo offers will help the organization to retain those customers and reduce revenue loss.
* The company can also internally study any possible operational causes and improve its product offerings.
* Proactive actions will prevent the loss of revenue for the company and will improve / retain the market share among the industry peers in terms of the number of active subscribers.

**OBJECTIVE**

* The objective is to predict to a high accuracy, in advance the customers who may attrite from the existing service provider in near future.
* Analyze using standard SEMMA (Sample, Explore, Modify, Model and Assess) approach and choose the best model based on the lowest Dollar ($) cost of misclassification.
* Recommend product strategies to business team based on analysis of product offerings that will help in retaining the customer based on available data.

**DATA DESCRIPTION**

* Data consists of 7043 fictional customers who belong to various demographics (single; with dependents; senior citizen) and subscribe to different products offerings (internet service; phone line; streaming TV; streaming movies; online security) from a telecom company located in one of the US states.
* Independent variables: 17 Categorical and 3 Continuous
* Dependent Target variable: “Churn”
* Churn Rate (Baseline) is 26.5%
* Dataset source: https://raw.githubusercontent.com/dsrscientist/DSData/master/Telecom\_customer\_churn.csv

Churn is a one of the biggest problem in the telecom industry. Research has shown that the average monthly churn rate among the top 4 wireless carriers in the US is 1.9% - 2%.

First we imported the libraries

**import** numpy **as** np *# linear algebra*

**import** pandas **as** pd

**import** seaborn **as** sns

**import** matplotlib.ticker **as** mtick

**import** matplotlib.pyplot **as** plt

sns**.**set(style **=** 'white')

After importing the libraries we had take the data

telecom\_cust **=** pd**.**read\_csv('https://raw.githubusercontent.com/dsrscientist/DSData/master/Telecom\_customer\_churn.csv')

telecom\_cust**.**head()

telecom\_cust**.**columns**.**values

Let's explore the data to see if there are any missing values.

*# Checking the data types of all the columns*

telecom\_cust**.**dtypes

*# Converting Total Charges to a numerical data type.*

telecom\_cust**.**TotalCharges **=** pd**.**to\_numeric(telecom\_cust**.**TotalCharges, errors**=**'coerce')

telecom\_cust**.**isnull()**.**sum()

After looking at the above output, we can say that there are 11 missing values for Total Charges. Let us replace remove these 11 rows from our data set

*#Removing missing values*

telecom\_cust**.**dropna(inplace **=** **True**)

*#Remove customer IDs from the data set*

df2 **=** telecom\_cust**.**iloc[:,1:]

*#Convertin the predictor variable in a binary numeric variable*

df2['Churn']**.**replace(to\_replace**=**'Yes', value**=**1, inplace**=True**)

df2['Churn']**.**replace(to\_replace**=**'No', value**=**0, inplace**=True**)

*#Let's convert all the categorical variables into dummy variables*

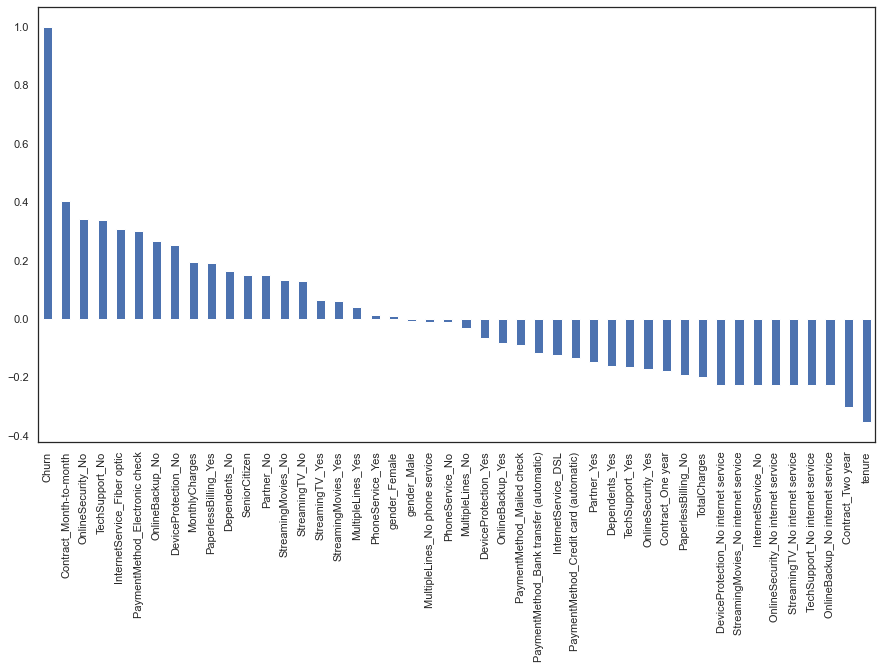
df\_dummies **=** pd**.**get\_dummies(df2)

df\_dummies**.**head()

*#Get Correlation of "Churn" with other variables:*

plt**.**figure(figsize**=**(15,8))

df\_dummies**.**corr()['Churn']**.**sort\_values(ascending **=** **False**)**.**plot(kind**=**'bar')



**DATA EXPLORATION**

Let us first start with exploring our data set, to better understand the patterns in the data and potentially form some hypothesis. First we will look at the distribution of individual variables and then slice and dice our data for any interesting trends.

A.) Demographics - Let us first understand the gender, age range, patner and dependent status of the customers

1.Gender Distribution - About half of the customers in our data set are male while the other half are female

colors **=** ['#4D3425','#E4512B']

ax **=** (telecom\_cust['gender']**.**value\_counts()**\***100.0 **/**len(telecom\_cust))**.**plot(kind**=**'bar',

stacked **=** **True**,

rot **=** 0,

color **=** colors)

ax**.**yaxis**.**set\_major\_formatter(mtick**.**PercentFormatter())

ax**.**set\_ylabel('% Customers')

ax**.**set\_xlabel('Gender')

ax**.**set\_ylabel('% Customers')

ax**.**set\_title('Gender Distribution')

*# create a list to collect the plt.patches data*

totals **=** []

*# find the values and append to list*

**for** i **in** ax**.**patches:

totals**.**append(i**.**get\_width())

*# set individual bar lables using above list*

total **=** sum(totals)

**for** i **in** ax**.**patches:

*# get\_width pulls left or right; get\_y pushes up or down*

ax**.**text(i**.**get\_x()**+**.15, i**.**get\_height()**-**3.5, \

str(round((i**.**get\_height()**/**total), 1))**+**'%',

fontsize**=**12,

color**=**'white',

weight **=** 'bold')

* Variables, “Tenure” and “MonthlyCharges”, both are positively corelated to “TotalCharges” and can be identified approximately as “TotalCharges = Tenure x MonthlyCharges”.
* In the scatterplot matrix, red dots represent the records which have churn as “no” and blue dots represent records with churn as “yes”.
* **Missing Values:** “Total Charges” has 11 missing values.
* **Outliers:** There are no outliers in the dataset.
* Data is partitioned into 60% Training; 20% Validation and 20% Test using formula random method.
* Month to month contracts, absence of online security and tech support seem to be positively correlated with churn. While, tenure, two year contracts seem to be negatively correlated with churn.
* Interestingly, services such as Online security, streaming TV, online backup, tech support, etc. without internet connection seem to be negatively related to churn.
* We will explore the patterns for the above correlations below before we delve into modelling and identifying the important variables.

1.% Senior Citizens - There are only 16% of the customers who are senior citizens. Thus most of our customers in the data are younger people.

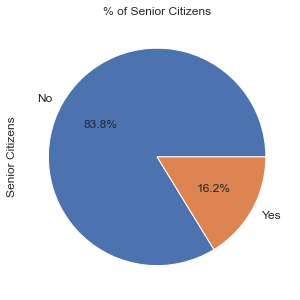
ax **=** (telecom\_cust['SeniorCitizen']**.**value\_counts()**\***100.0 **/**len(telecom\_cust))\

**.**plot**.**pie(autopct**=**'%.1f%%', labels **=** ['No', 'Yes'],figsize **=**(5,5), fontsize **=** 12 )

ax**.**yaxis**.**set\_major\_formatter(mtick**.**PercentFormatter())

ax**.**set\_ylabel('Senior Citizens',fontsize **=** 12)

ax**.**set\_title('% of Senior Citizens', fontsize **=** 12)



Partner and dependent status - About 50% of the customers have a partner, while only 30% of the total customers have dependents.

df2 **=** pd**.**melt(telecom\_cust, id\_vars**=**['customerID'], value\_vars**=**['Dependents','Partner'])

df3 **=** df2**.**groupby(['variable','value'])**.**count()**.**unstack()

df3 **=** df3**\***100**/**len(telecom\_cust)

colors **=** ['#4D3425','#E4512B']

ax **=** df3**.**loc[:,'customerID']**.**plot**.**bar(stacked**=True**, color**=**colors,

figsize**=**(8,6),rot **=** 0,

width **=** 0.2)

ax**.**yaxis**.**set\_major\_formatter(mtick**.**PercentFormatter())

ax**.**set\_ylabel('% Customers',size **=** 14)

ax**.**set\_xlabel('')

ax**.**set\_title('% Customers with dependents and partners',size **=** 14)

ax**.**legend(loc **=** 'center',prop**=**{'size':14})

**for** p **in** ax**.**patches:

width, height **=** p**.**get\_width(), p**.**get\_height()

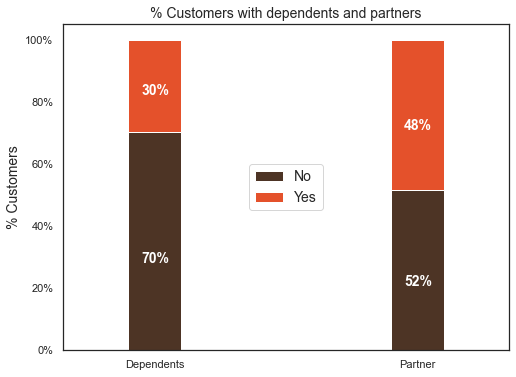
x, y **=** p**.**get\_xy()

ax**.**annotate('{:.0f}%'**.**format(height), (p**.**get\_x()**+**.25**\***width, p**.**get\_y()**+**.4**\***height),

color **=** 'white',

weight **=** 'bold',

size **=** 14)



What would be interesting is to look at the % of customers, who have partners, also have dependents. We will explore this next.

Interestingly, among the customers who have a partner, only about half of them also have a dependent, while other half do not have any independents. Additionally, as expected, among the customers who do not have any partner, a majority (80%) of them do not have any dependents .

colors **=** ['#4D3425','#E4512B']

partner\_dependents **=** telecom\_cust**.**groupby(['Partner','Dependents'])**.**size()**.**unstack()

ax **=** (partner\_dependents**.**T**\***100.0 **/** partner\_dependents**.**T**.**sum())**.**T**.**plot(kind**=**'bar',

width **=** 0.2,

stacked **=** **True**,

rot **=** 0,

figsize **=** (8,6),

color **=** colors)

ax**.**yaxis**.**set\_major\_formatter(mtick**.**PercentFormatter())

ax**.**legend(loc**=**'center',prop**=**{'size':14},title **=** 'Dependents',fontsize **=**14)

ax**.**set\_ylabel('% Customers',size **=** 14)

ax**.**set\_title('% Customers with/without dependents based on whether they have a partner',size **=** 14)

ax**.**xaxis**.**label**.**set\_size(14)

*# Code to add the data labels on the stacked bar chart*

**for** p **in** ax**.**patches:

width, height **=** p**.**get\_width(), p**.**get\_height()

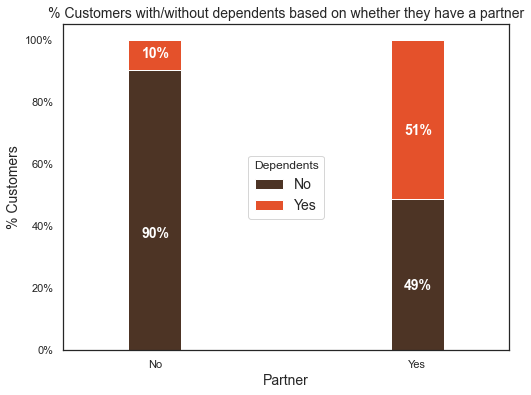
x, y **=** p**.**get\_xy()

ax**.**annotate('{:.0f}%'**.**format(height), (p**.**get\_x()**+**.25**\***width, p**.**get\_y()**+**.4**\***height),

color **=** 'white',

weight **=** 'bold',

size **=** 14)



I also looked at any differences between the % of customers with/without dependents and partners by gender. There is no difference in their distribution by gender. Additionally, there is no difference in senior citizen status by gender.

B.) Customer Account Information: Let u now look at the tenure, contract

1. Tenure: After looking at the below histogram we can see that a lot of customers have been with the telecom company for just a month, while quite a many are there for about 72 months. This could be potentially because different customers have different contracts. Thus based on the contract they are into it could be more/less easier for the customers to stay/leave the telecom company.

ax **=** sns**.**distplot(telecom\_cust['tenure'], hist**=True**, kde**=False**,

bins**=**int(180**/**5), color **=** 'darkblue',

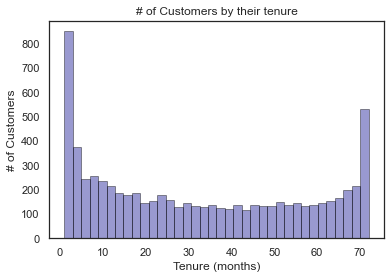
hist\_kws**=**{'edgecolor':'black'},

kde\_kws**=**{'linewidth': 4})

ax**.**set\_ylabel('# of Customers')

ax**.**set\_xlabel('Tenure (months)')

ax**.**set\_title('# of Customers by their tenure')

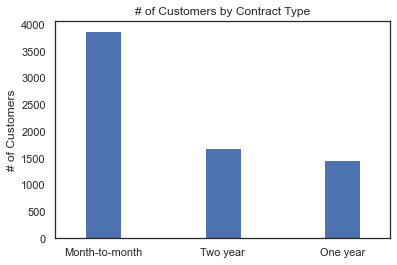


1. Contracts: To understand the above graph, lets first look at the # of customers by different contracts.

ax **=** telecom\_cust['Contract']**.**value\_counts()**.**plot(kind **=** 'bar',rot **=** 0, width **=** 0.3)

ax**.**set\_ylabel('# of Customers')

ax**.**set\_title('# of Customers by Contract Type')



Let us now look at the distribution of various services used by customers

telecom\_cust**.**columns**.**values

services **=** ['PhoneService','MultipleLines','InternetService','OnlineSecurity',

'OnlineBackup','DeviceProtection','TechSupport','StreamingTV','StreamingMovies']

fig, axes **=** plt**.**subplots(nrows **=** 3,ncols **=** 3,figsize **=** (15,12))

**for** i, item **in** enumerate(services):

**if** i **<** 3:

ax **=** telecom\_cust[item]**.**value\_counts()**.**plot(kind **=** 'bar',ax**=**axes[i,0],rot **=** 0)

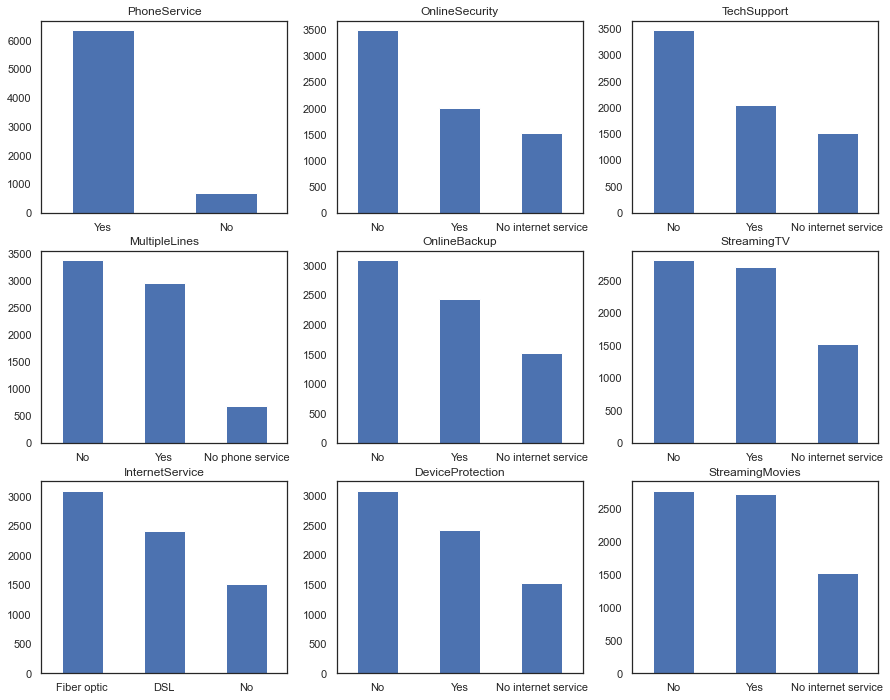
**elif** i **>=**3 **and** i **<** 6:

ax **=** telecom\_cust[item]**.**value\_counts()**.**plot(kind **=** 'bar',ax**=**axes[i**-**3,1],rot **=** 0)

**elif** i **<** 9:

ax **=** telecom\_cust[item]**.**value\_counts()**.**plot(kind **=** 'bar',ax**=**axes[i**-**6,2],rot **=** 0)

ax**.**set\_title(item)

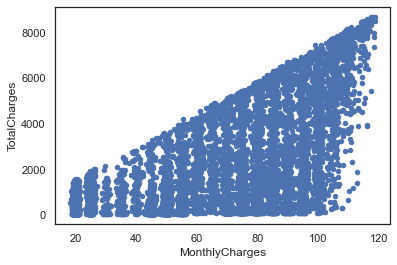


Now let's take a quick look at the relation between monthly and total charges

We will observe that the total charges increases as the monthly bill for a customer increases.

telecom\_cust[['MonthlyCharges', 'TotalCharges']]**.**plot**.**scatter(x **=** 'MonthlyCharges',

y**=**'TotalCharges')



Finally, let's take a look at out predictor variable (Churn) and understand its interaction with other important variables as was found out in the correlation plot.

1.Lets first look at the churn rate in our data

colors **=** ['#4D3425','#E4512B']

ax **=** (telecom\_cust['Churn']**.**value\_counts()**\***100.0 **/**len(telecom\_cust))**.**plot(kind**=**'bar',

stacked **=** **True**,

rot **=** 0,

color **=** colors,

figsize **=** (8,6))

ax**.**yaxis**.**set\_major\_formatter(mtick**.**PercentFormatter())

ax**.**set\_ylabel('% Customers',size **=** 14)

ax**.**set\_xlabel('Churn',size **=** 14)

ax**.**set\_title('Churn Rate', size **=** 14)

*# create a list to collect the plt.patches data*

totals **=** []

*# find the values and append to list*

**for** i **in** ax**.**patches:

totals**.**append(i**.**get\_width())

*# set individual bar lables using above list*

total **=** sum(totals)

**for** i **in** ax**.**patches:

*# get\_width pulls left or right; get\_y pushes up or down*

ax**.**text(i**.**get\_x()**+**.15, i**.**get\_height()**-**4.0, \

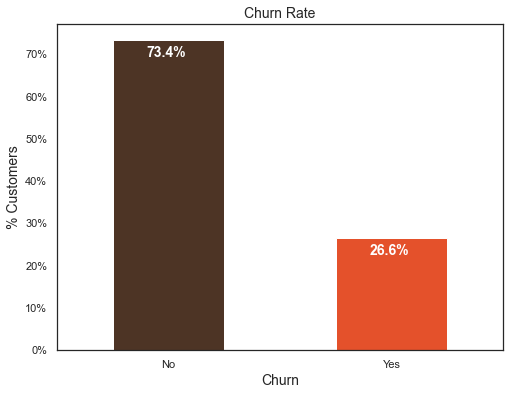
str(round((i**.**get\_height()**/**total), 1))**+**'%',

fontsize**=**12,

color**=**'white',

weight **=** 'bold',

size **=** 14)

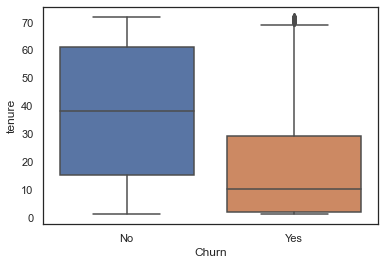


In our data, 74% of the customers do not churn. Clearly the data is skewed as we would expect a large majority of the customers to not churn. This is important to keep in mind for our modelling as skeweness could lead to a lot of false negatives. We will see in the modelling section on how to avoid skewness in the data.

Lets now explore the churn rate by tenure, seniority, contract type, monthly charges and total charges to see how it varies by these variables.

i.) Churn vs Tenure: As we can see form the below plot, the customers who do not churn, they tend to stay for a longer tenure with the telecom company.

sns**.**boxplot(x **=** telecom\_cust**.**Churn, y **=** telecom\_cust**.**tenure)



ii.) Churn by Contract Type: Similar to what we saw in the correlation plot, the customers who have a month to month contract have a very high churn rate.

colors **=** ['#4D3425','#E4512B']

contract\_churn **=** telecom\_cust**.**groupby(['Contract','Churn'])**.**size()**.**unstack()

ax **=** (contract\_churn**.**T**\***100.0 **/** contract\_churn**.**T**.**sum())**.**T**.**plot(kind**=**'bar',

width **=** 0.3,

stacked **=** **True**,

rot **=** 0,

figsize **=** (10,6),

color **=** colors)

ax**.**yaxis**.**set\_major\_formatter(mtick**.**PercentFormatter())

ax**.**legend(loc**=**'best',prop**=**{'size':14},title **=** 'Churn')

ax**.**set\_ylabel('% Customers',size **=** 14)

ax**.**set\_title('Churn by Contract Type',size **=** 14)

*# Code to add the data labels on the stacked bar chart*

**for** p **in** ax**.**patches:

width, height **=** p**.**get\_width(), p**.**get\_height()

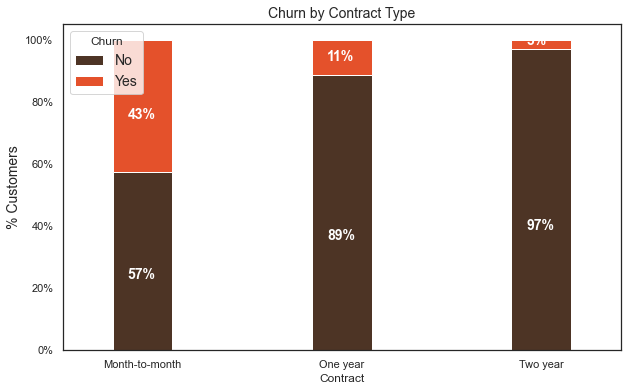
x, y **=** p**.**get\_xy()

ax**.**annotate('{:.0f}%'**.**format(height), (p**.**get\_x()**+**.25**\***width, p**.**get\_y()**+**.4**\***height),

color **=** 'white',

weight **=** 'bold',

size **=** 14)



iii.) Churn by Seniority: Senior Citizens have almost double the churn rate than younger population.

colors **=** ['#4D3425','#E4512B']

seniority\_churn **=** telecom\_cust**.**groupby(['SeniorCitizen','Churn'])**.**size()**.**unstack()

ax **=** (seniority\_churn**.**T**\***100.0 **/** seniority\_churn**.**T**.**sum())**.**T**.**plot(kind**=**'bar',

width **=** 0.2,

stacked **=** **True**,

rot **=** 0,

figsize **=** (8,6),

color **=** colors)

ax**.**yaxis**.**set\_major\_formatter(mtick**.**PercentFormatter())

ax**.**legend(loc**=**'center',prop**=**{'size':14},title **=** 'Churn')

ax**.**set\_ylabel('% Customers')

ax**.**set\_title('Churn by Seniority Level',size **=** 14)

*# Code to add the data labels on the stacked bar chart*

**for** p **in** ax**.**patches:

width, height **=** p**.**get\_width(), p**.**get\_height()

x, y **=** p**.**get\_xy()

ax**.**annotate('{:.0f}%'**.**format(height), (p**.**get\_x()**+**.25**\***width, p**.**get\_y()**+**.4**\***height),

color **=** 'white',

weight **=** 'bold',size **=**14)

iv.) Churn by Monthly Charges: Higher % of customers churn when the monthly charges are high.

ax **=** sns**.**kdeplot(telecom\_cust**.**MonthlyCharges[(telecom\_cust["Churn"] **==** 'No') ],

color**=**"Red", shade **=** **True**)

ax **=** sns**.**kdeplot(telecom\_cust**.**MonthlyCharges[(telecom\_cust["Churn"] **==** 'Yes') ],

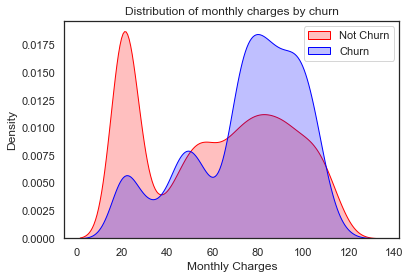
ax **=**ax, color**=**"Blue", shade**=** **True**)

ax**.**legend(["Not Churn","Churn"],loc**=**'upper right')

ax**.**set\_ylabel('Density')

ax**.**set\_xlabel('Monthly Charges')

ax**.**set\_title('Distribution of monthly charges by churn')



v.) Churn by Total Charges: It seems that there is higer churn when the total charges are lower.

ax **=** sns**.**kdeplot(telecom\_cust**.**TotalCharges[(telecom\_cust["Churn"] **==** 'No') ],

color**=**"Red", shade **=** **True**)

ax **=** sns**.**kdeplot(telecom\_cust**.**TotalCharges[(telecom\_cust["Churn"] **==** 'Yes') ],

ax **=**ax, color**=**"Blue", shade**=** **True**)

ax**.**legend(["Not Churn","Churn"],loc**=**'upper right')

ax**.**set\_ylabel('Density')

ax**.**set\_xlabel('Total Charges')

ax**.**set\_title('Distribution of total charges by churn')

**After going through the above EDA we will develop some predictive models and compare them.**

We will develop Logistic Regression, Random Forest, SVM, ADA Boost and XG Boost

1. Logistic Regression

*# We will use the data frame where we had created dummy variables*

y **=** df\_dummies['Churn']**.**values

X **=** df\_dummies**.**drop(columns **=** ['Churn'])

*# Scaling all the variables to a range of 0 to 1*

**from** sklearn.preprocessing **import** MinMaxScaler

features **=** X**.**columns**.**values

scaler **=** MinMaxScaler(feature\_range **=** (0,1))

scaler**.**fit(X)

X **=** pd**.**DataFrame(scaler**.**transform(X))

X**.**columns **=** features

It is important to scale the variables in logistic regression so that all of them are within a range of 0 to 1. This helped me improve the accuracy from 79.7% to 80.7%. Further, you will notice below that the importance of variables is also aligned with what we are seeing in Random Forest algorithm and the EDA we conducted above.

*# Create Train & Test Data*

**from** sklearn.model\_selection **import** train\_test\_split

X\_train, X\_test, y\_train, y\_test **=** train\_test\_split(X, y, test\_size**=**0.3, random\_state**=**101)

*# Running logistic regression model*

**from** sklearn.linear\_model **import** LogisticRegression

model **=** LogisticRegression()

result **=** model**.**fit(X\_train, y\_train)

**from** sklearn **import** metrics

prediction\_test **=** model**.**predict(X\_test)

*# Print the prediction accuracy*

print (metrics**.**accuracy\_score(y\_test, prediction\_test))

*# To get the weights of all the variables*

weights **=** pd**.**Series(model**.**coef\_[0],

index**=**X**.**columns**.**values)

print (weights**.**sort\_values(ascending **=** **False**)[:10]**.**plot(kind**=**'bar'))

**Observations**

We can see that some variables have a negative relation to our predicted variable (Churn), while some have positive relation. Negative relation means that likeliness of churn decreases with that variable. Let us summarize some of the interesting features below:

As we saw in our EDA, having a 2 month contract reduces chances of churn. 2 month contract along with tenure have the most negative relation with Churn as predicted by logistic regressions Having DSL internet service also reduces the proability of Churn Lastly, total charges, monthly contracts, fibre optic internet services and seniority can lead to higher churn rates. This is interesting because although fibre optic services are faster, customers are likely to churn because of it. I think we need to explore more to better understad why this is happening. Any hypothesis on the above would be really helpful!

**CONCLUSION**

* Significant variables impacting “Churn”: Type & Tenure of Contract
* Churn is observed to be high for customers:
  + Without dependents
  + With high cost Phone Services
  + Having single line service (no combo services)
* Recommendation to Business Team for retaining Customer
  + Targeted Customer Promotion
  + Promote Long Term contract
  + Market more products as Combo (multi) service offerings