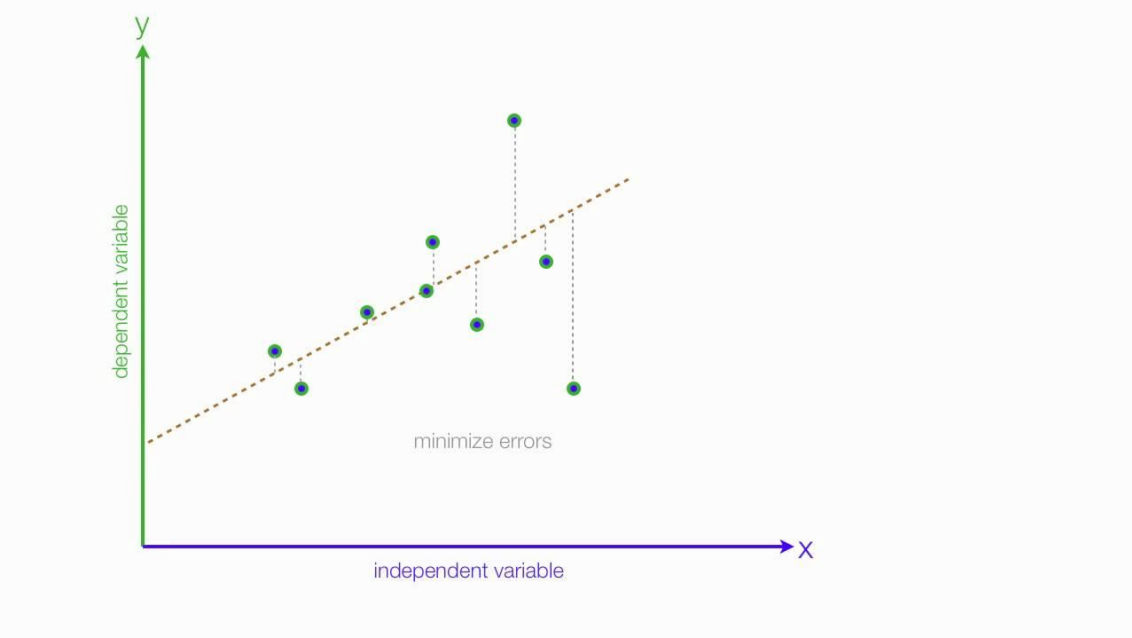
TELCO CUSTOMER ANALYTICS

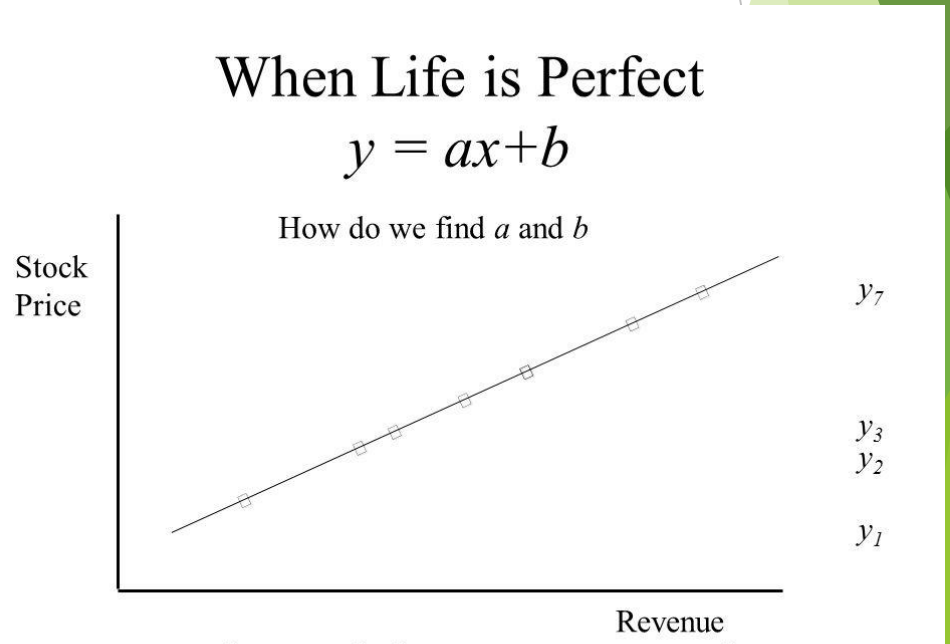
By Vinh Vu

LinkedIn: https://www.linkedin.com/in/vinh-v-0b9059163/

Machine learning

- **A regression problem** is when the output variable is a real or continuous value, such as “salary” or “weight”. Many different models can be used, the simplest is the linear regression. It tries to fit data with the best hyper-plane which goes through the points.





About the project:

- My client, Telco Company, is a telephone and internet service provider with over 5000 customers. In order to grow and maintain profitability, it's essential that they learn how to maintain a dedicated customer base and reduce churn. Based on my analysis, Telco can:
  - Identify customers that are likely to churn and reach out to them to try to stop them from churning via special offers targeted to their needs
  - Focus marketing on customers that are more likely to be long term customers
  - Modify their services to improve the likelihood customers will stay longer term

Data set

Attribute	Description
CustomerID	Customer ID
Gender	Customer gender (female, male)
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)

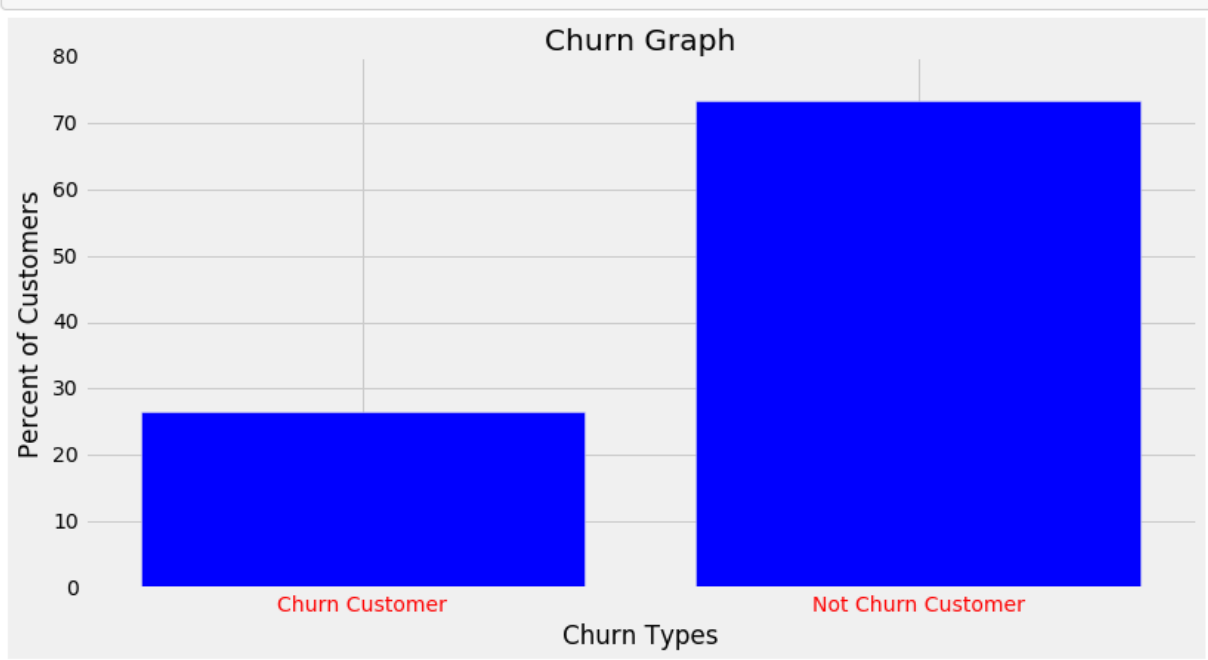
Constructing the Data:

- Convert columns with yes/ no to 1/0
- When a column has multiple values, I use get_dummy turning a column to multiple columns

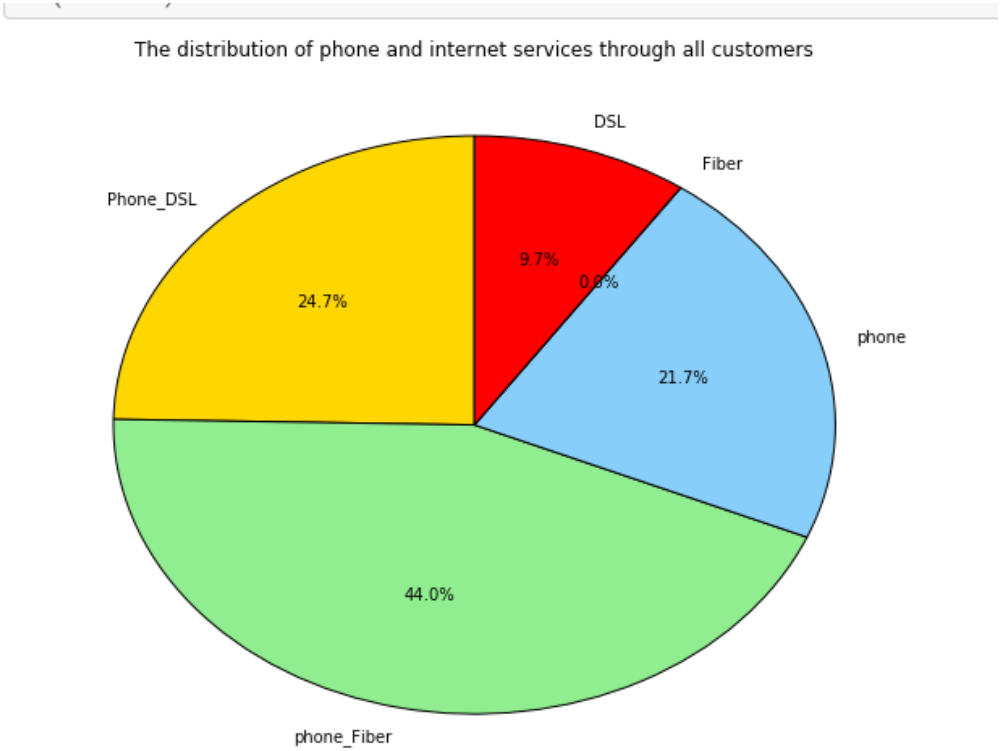
Customer ID	Payment Type
4562	automatic desposit
4563	cash
4564	check
4566	automatic desposit
4567	cash
4568	check

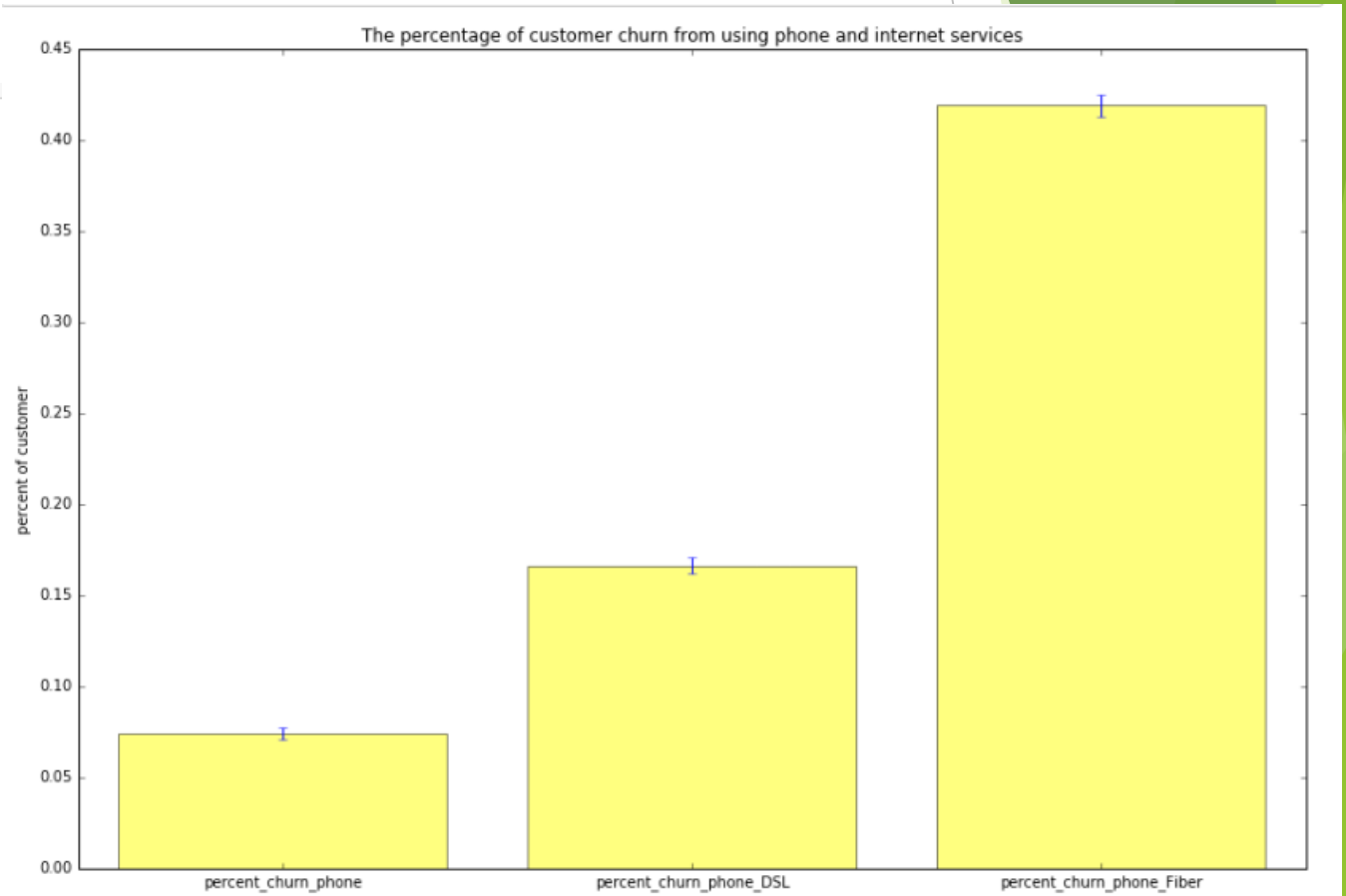
Customer ID	automatic desposit	cash	check
4562	1	0	0
4563	0	1	0
4564	0	0	1
4566	1	0	0
4567	0	1	0
4568	0	0	1

Exploring the Data

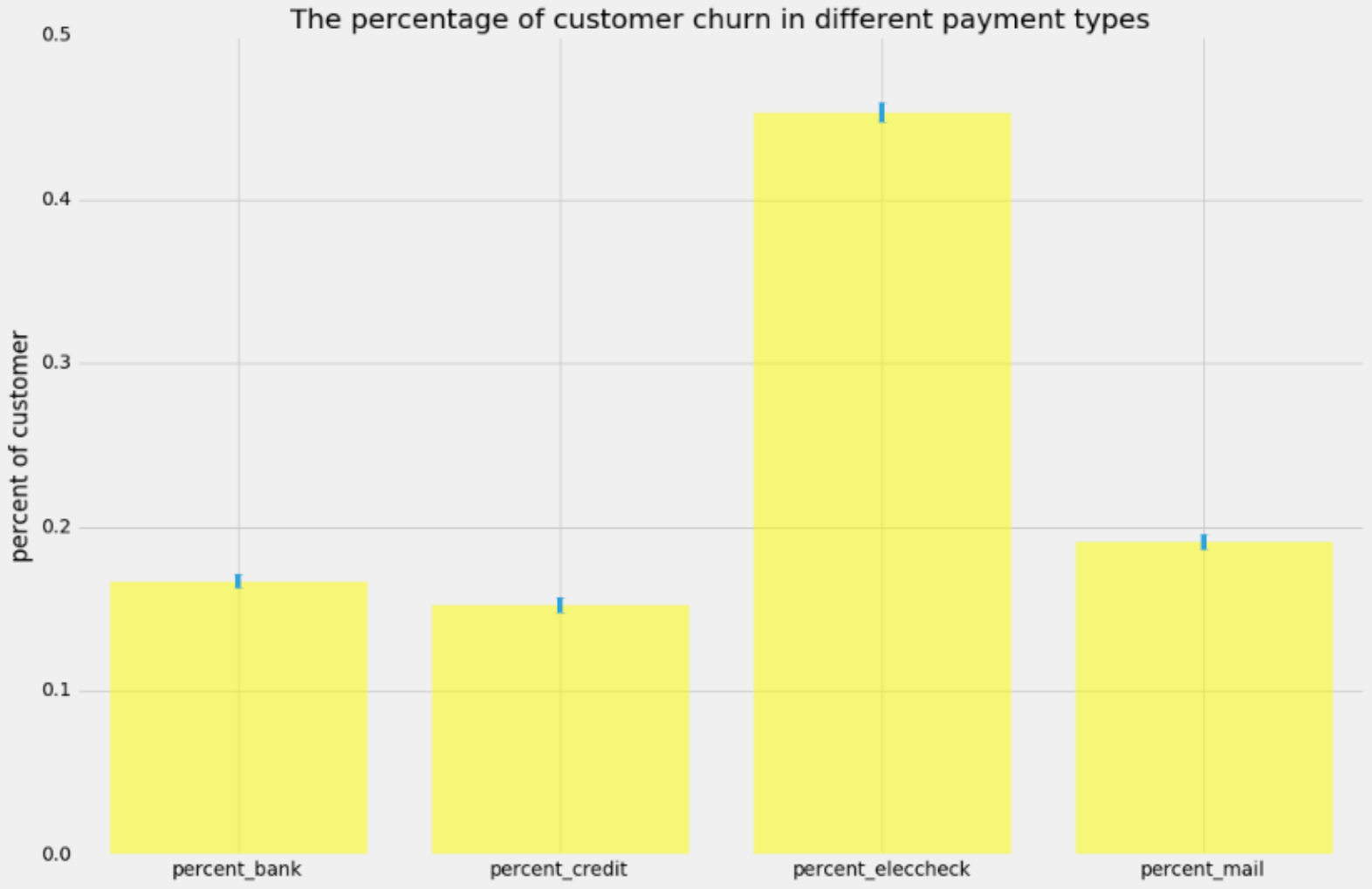


Services:

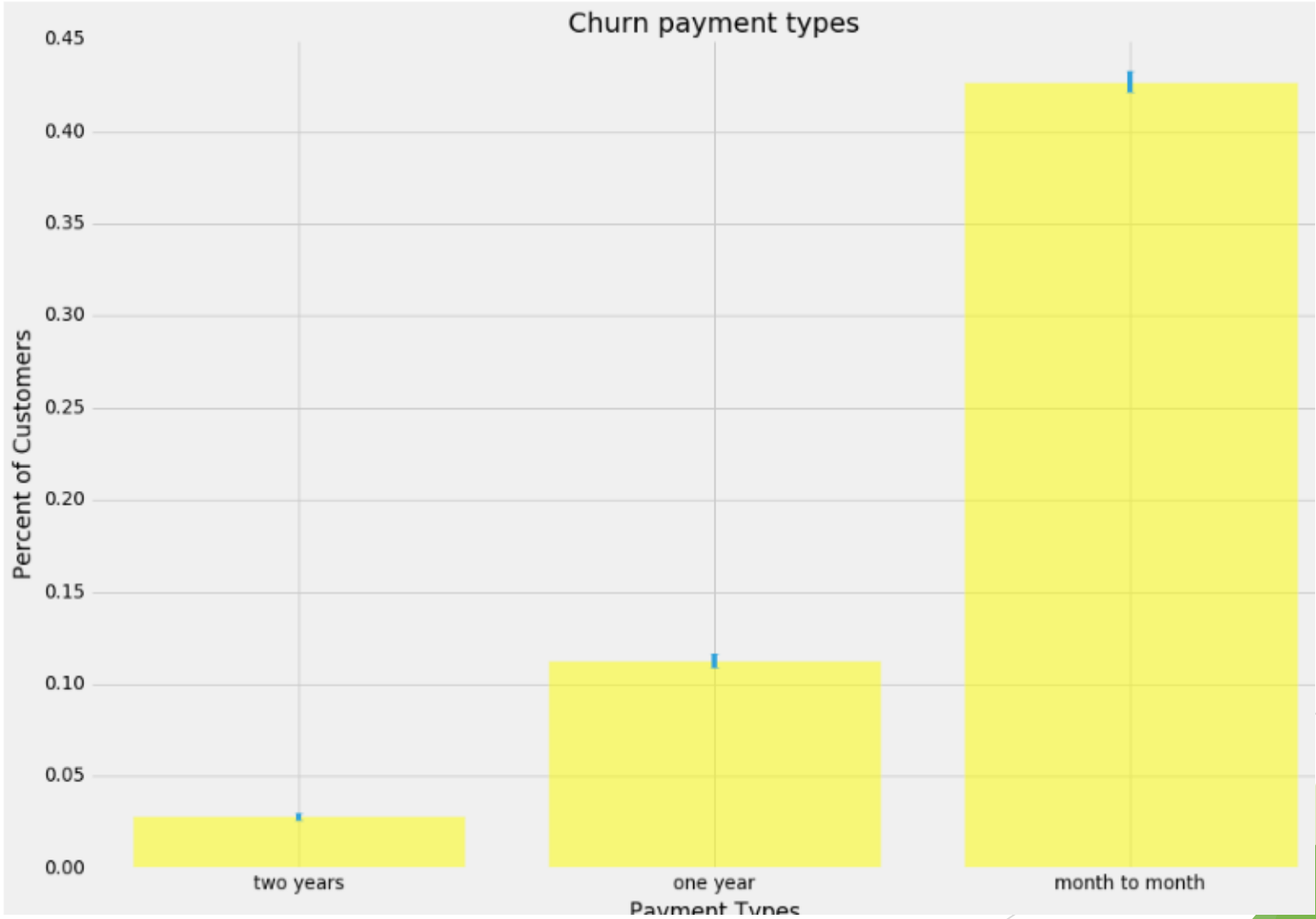

The distribution of phone and internet services through all customers
DSL
Fiber
Phone_DSL
9.7%
0.6%
24.7%
phone
21.7%
44.0%
phone_Fiber


The percentage of customer churn from using phone and internet services
percent of customer
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
percent_churn_phone
percent_churn_phone_DSL
percent_churn_phone_Fiber

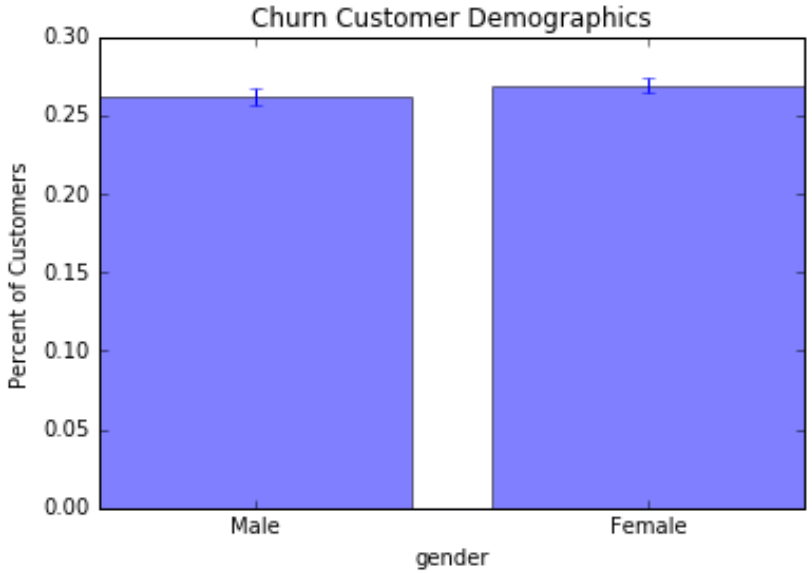
Payment types:

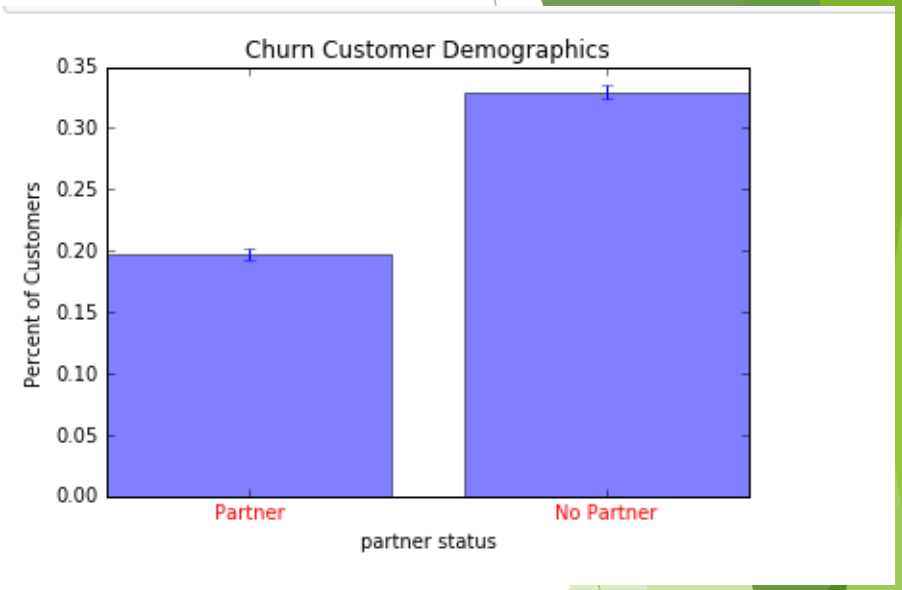


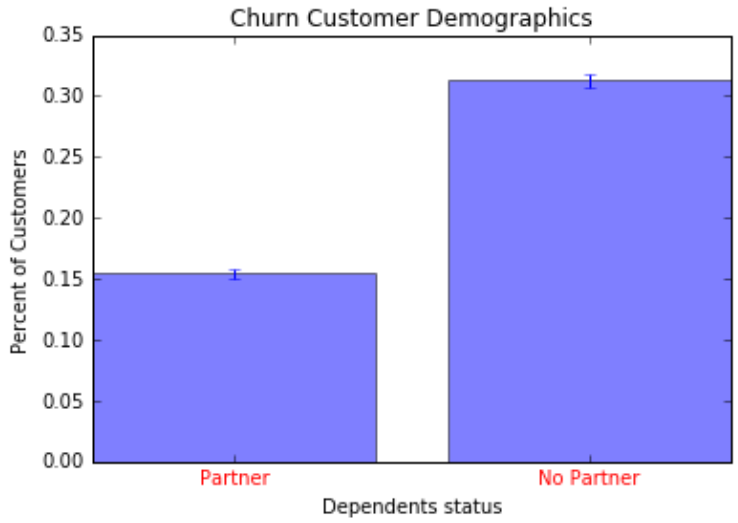
Payment types:



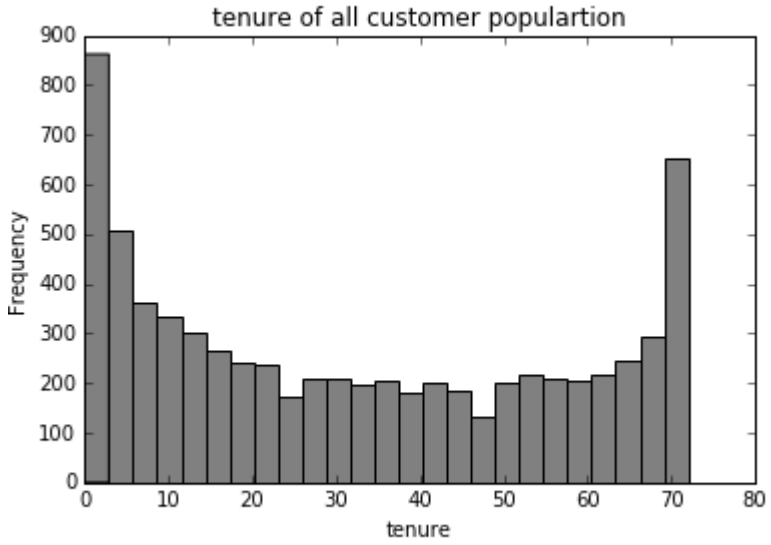
Demographic:

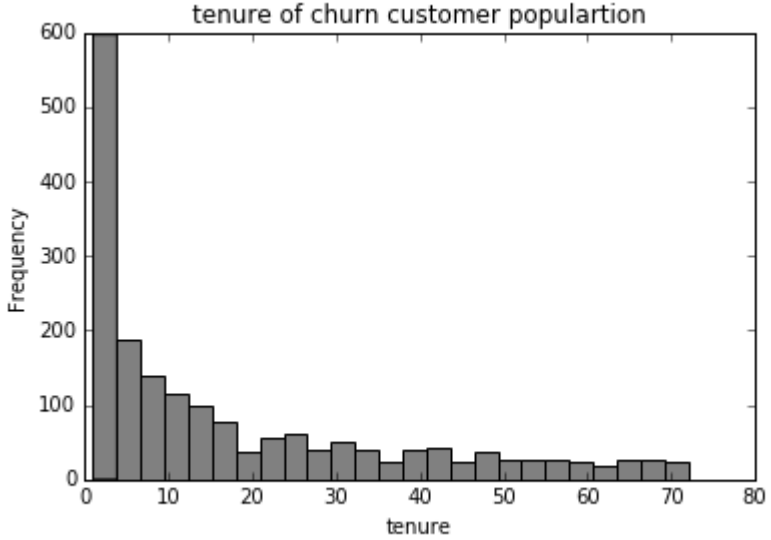


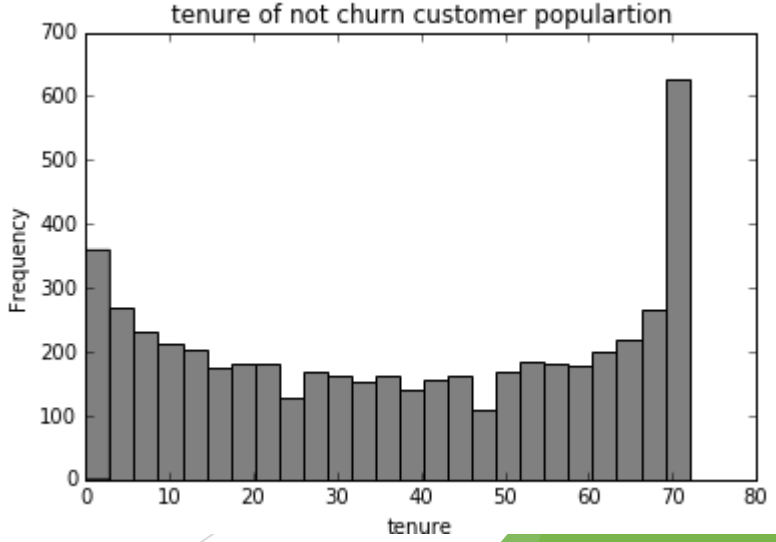




Tenure:


tenure of all customer popularion
Frequency
tenure
0
100
200
300
400
500
600
700
800
900
0
10
20
30
40
50
60
70
80


tenure of churn customer popularion
Frequency
tenure
0
100
200
300
400
500
600
0
10
20
30
40
50
60
70
80


tenure of not churn customer popularion
Frequency
tenure
0
100
200
300
400
500
600
700
0
10
20
30
40
50
60
70
80

Machine learning:

Due to skewed distribution of customers who churn and don't churn, the data set is an imbalanced problem. Imbalanced data typically refers to a problem with classification problems where the classes are not represented equally. In order to face imbalanced class, we measure model performance through different errors and concepts; such as Precision, recall. Due to this project, it is a binary class problem. Precision and recall are very important , because Precision is sort of like accuracy but it looks only at the data you predicted positive (in this example you're only looking at data where you predict a churn) and the recall is also sort of like accuracy but it looks only at the data that is "relevant" in some way.

Due this project is classification problem, I will use grid search in order to find the best parameter. Then it is used other models like logistic regression , random forest and decision tree in order to find the best fit model.

Logistic regression:

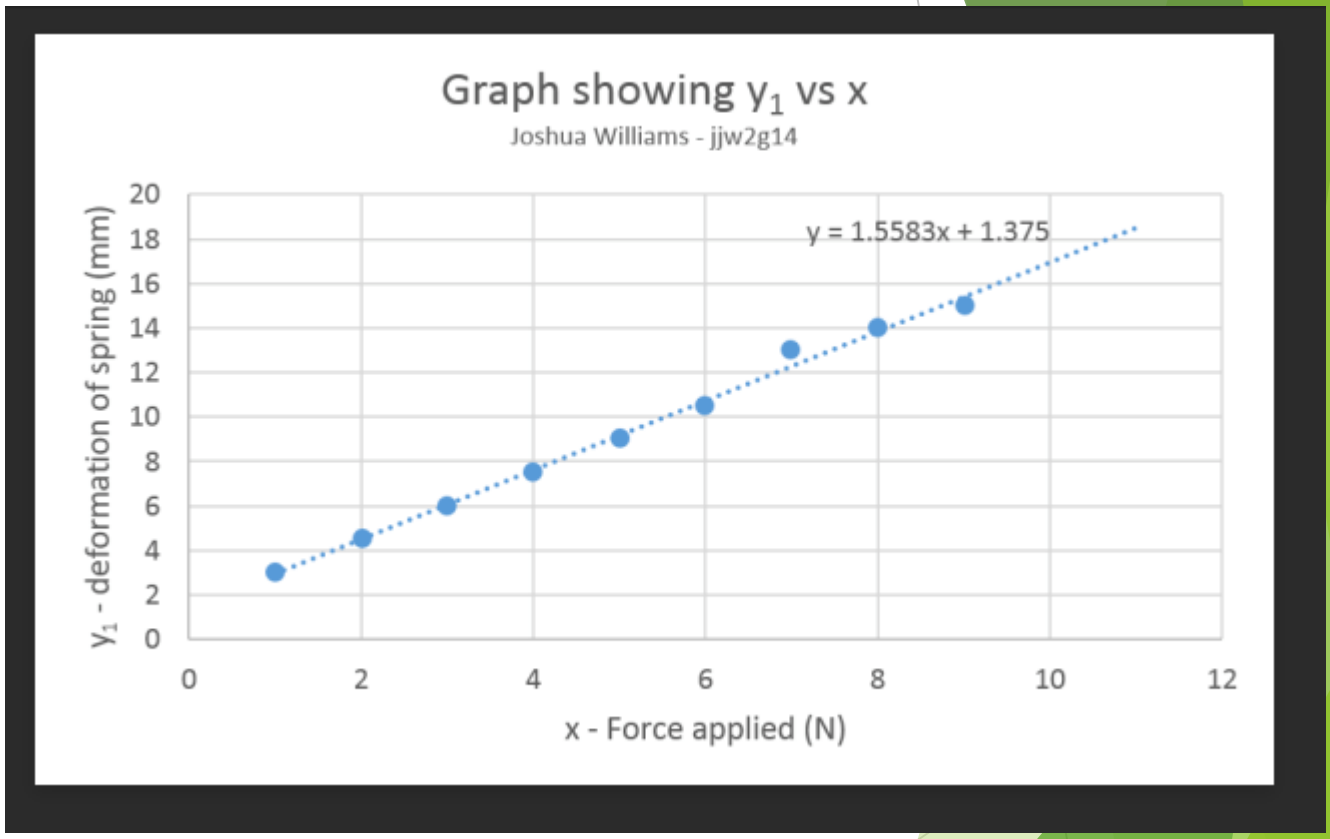
Linear Regression:

- **Simple:**

$$y = b_0 + b_1 * x$$

- **Multiple:**

$$y = b_0 + b_1 * x_1 + \ldots + b_n * x_n$$



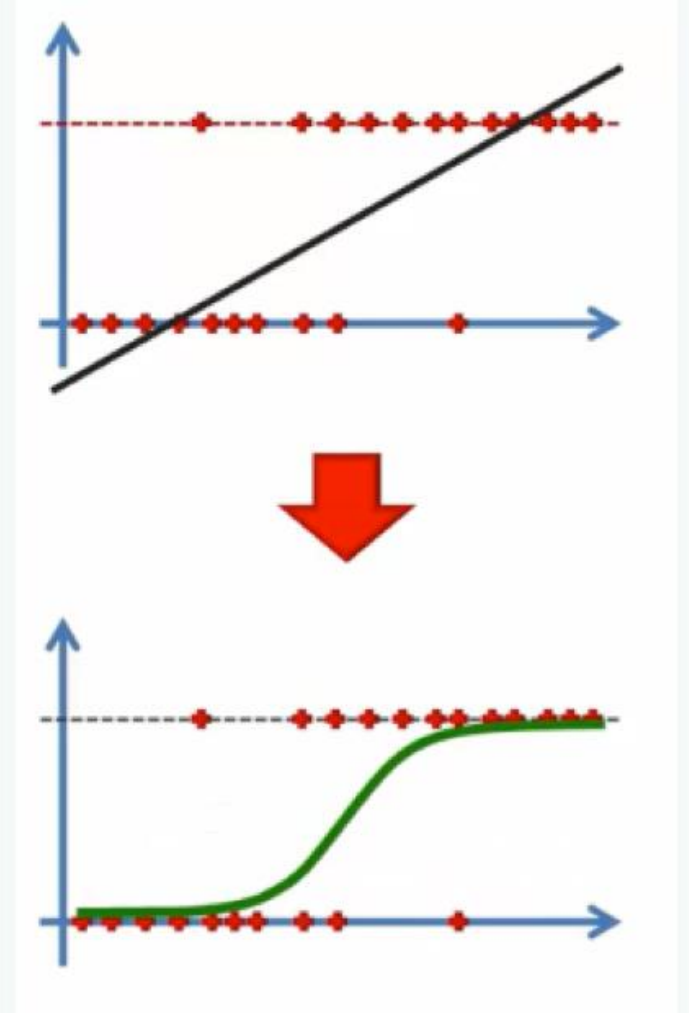
Logistic regression:

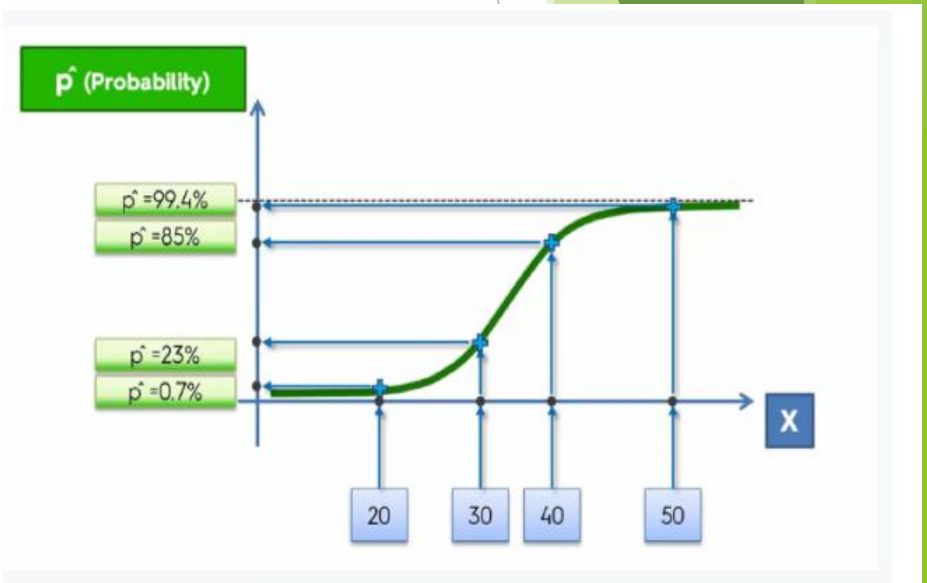
$$y = b_0 + b_1 * x$$

Sigmoid Function

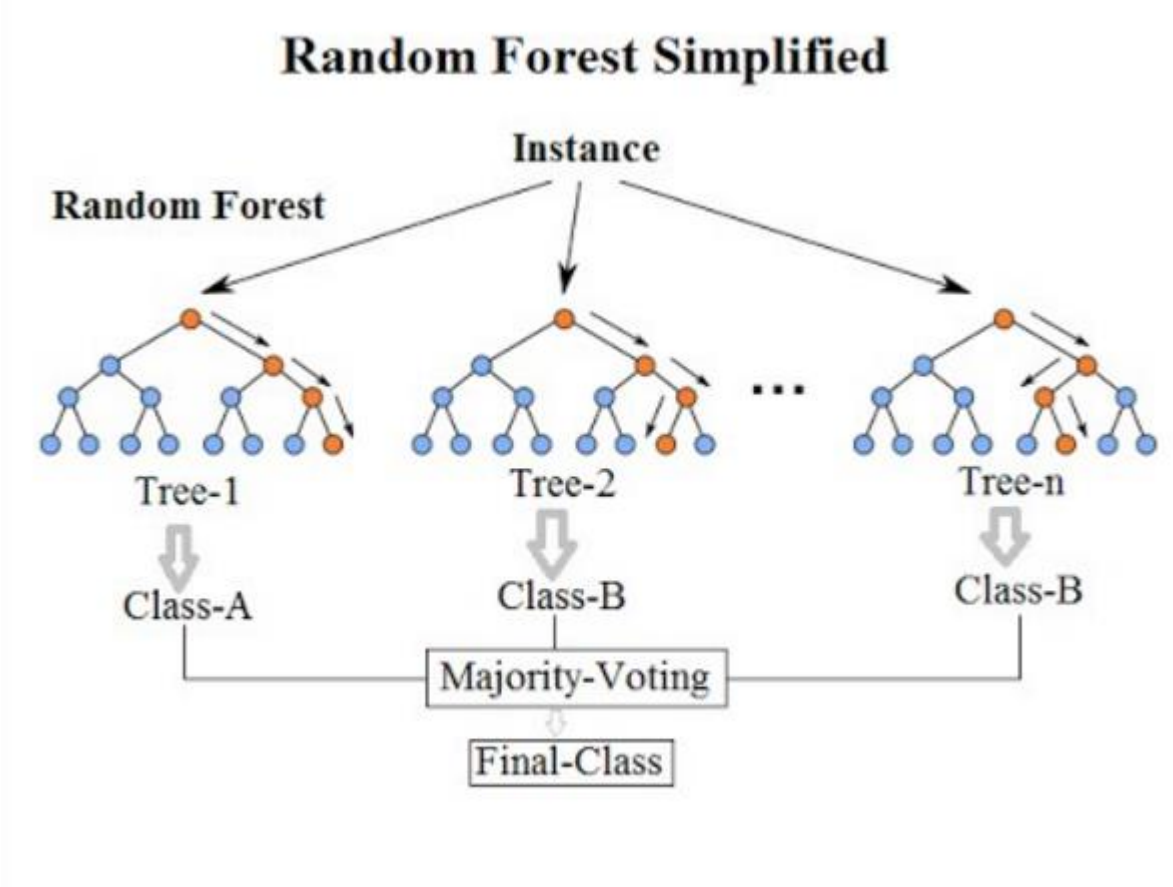
$$p = \frac{1}{1 + e^{-y}}$$

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 * x$$

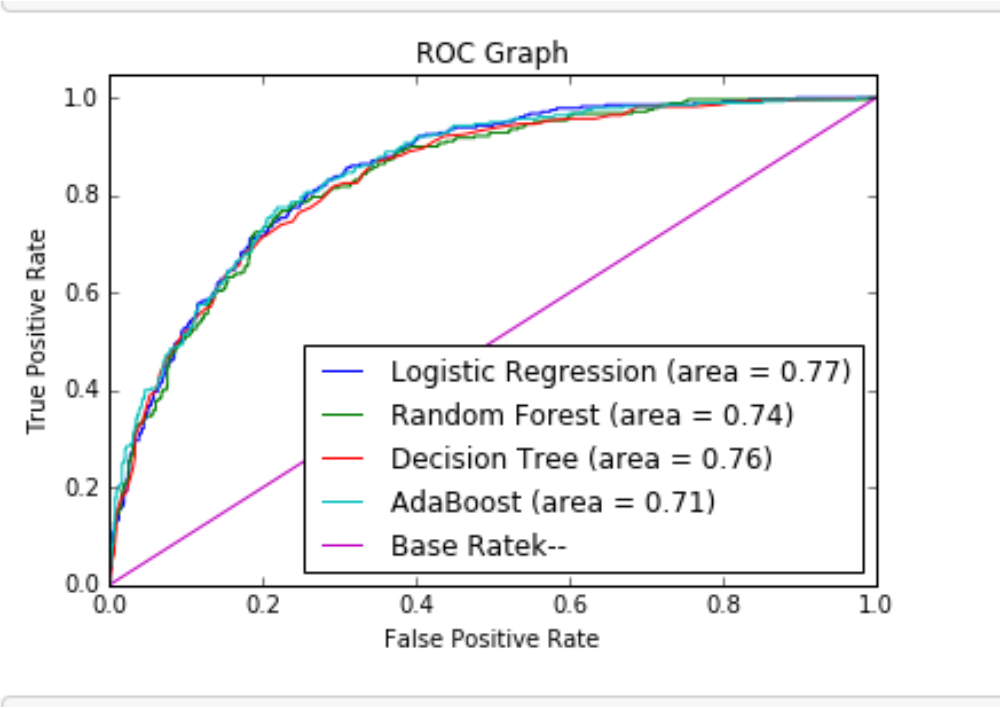




Random forest:

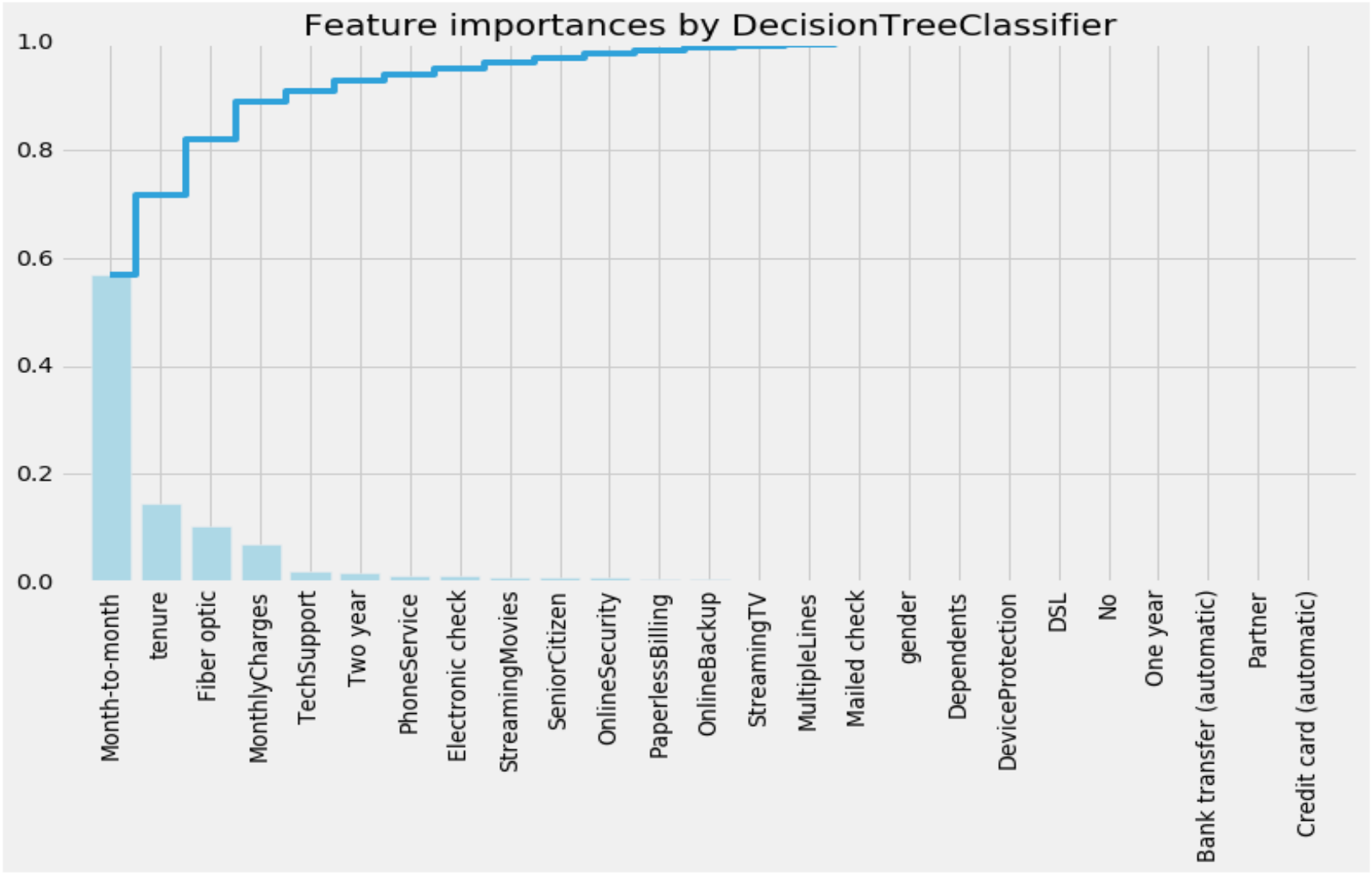


ROC curve:



In a Receiver Operating Characteristic (ROC) curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. A test with perfect discrimination (no overlap in the two distributions) has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test (Zweig & Campbell, 1993). According to our ROC graph, the logistic regression has a best fit model for this problem because AUC =0.77

Feature important:



Probability of customer churn:

Dep. Variable:	Churn	**No. Observations:**	7043
Model:	Logit	**Df Residuals:**	7039
Method:	MLE	**Df Model:**	3
Date:	Fri, 04 Jan 2019	**Pseudo R-squ.:**	0.2406
Time:	09:25:01	**Log-Likelihood:**	-3094.4
converged:	True	**LL-Null:**	-4075.1
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
Month-to-month	1.2010	0.095	12.590	0.000	1.014	1.388
tenure	-0.0308	0.002	-16.375	0.000	-0.034	-0.027
Fiber optic	1.5123	0.068	22.227	0.000	1.379	1.646
int	-1.8196	0.105	-17.260	0.000	-2.026	-1.613

$$\text{logit}\left[\theta(x)\right] = \log\left[\frac{\theta(x)}{1-\theta(x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

The equation above shows the relationship between the dependent/independent:

(θ(x)) - Dependent Variable (Outcome)
(xi) - Independent variables or predictor of event
(α) - is the constant of the equation
(β) - is the coefficient of the predictor variables or weights

Hypothetical example:

if the customer has month to month payment, fiber internet, and the tenure is 3 months:

```
y1=coef[3]+coef[0]*1+coef[1]*3+coef[2]*1
print(y1)
```

```
0.801285477303
```

Conclusion:

In summary, this is what we know about why customers churn:
1.Customer are most likely to churn if they are the month to month payment type.
2.Customer is likely to churn in first ten months.
3.Customers is likely to leave when they use Fiber optic service.
4.there is 0% customers use fiber only, customers are likely to cancel services when they use phone and fiber.
5.Customers with no partner or dependent are more likely to churn
6.month to month, tenure, and fiber optic are the three most significant features in determining churn.

Using the Logistic Regression model built to classify churn, wee can determine the probability each customer will churn. Telco can use this knowledge both to offer an array of products and services more likely to retain customers, and also use this knowledge to allocate resources to provide services specifically aimed at customers likely to churn so they won't leave. One primary weakness in this analysis, is that we cannot predict the exact timeframe in which a customer would leave, a next step in this analysis would be to take this into account using a time series based approach.