Machine learning : classification

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.

minimize errors
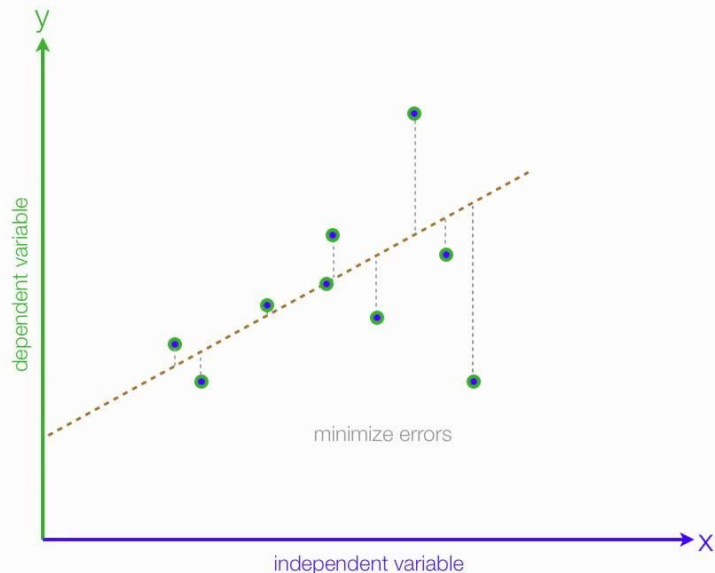
When Life is Perfect
$$y = ax+b$$
How do we find a and b

Stock Price

y_7

y_3
y_2

y_1

Revenue
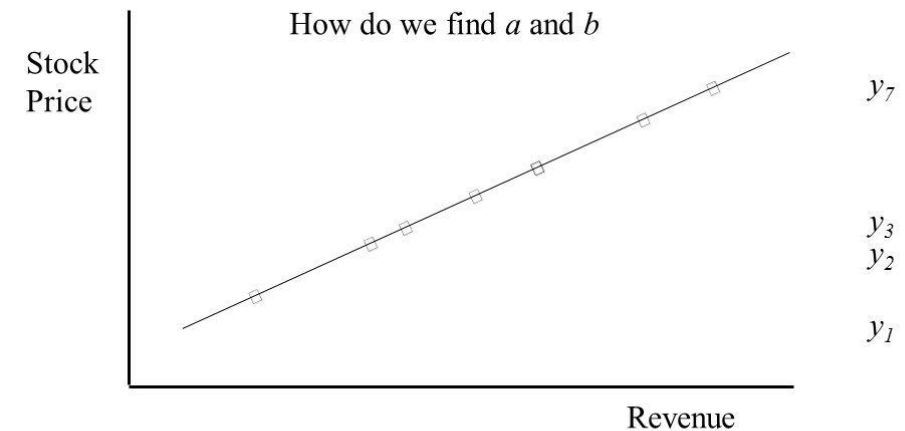
▶ **A classification problem** is when the output variable is a category, such as "red" or "blue" or "disease" and "no disease". A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes.

Predicted Class	True Outcome : Patients have Disease A	
	Positive (Patients have disease A)	Negative (Patients do not have disease A)
Positive (Patients have disease A)	True Positives	False Positives (Patients wrongly identified to have disease A)
Negative (Patients do not have disease A)	False Negatives (Patients have been left out from treatment for Disease)	True Negatives

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:
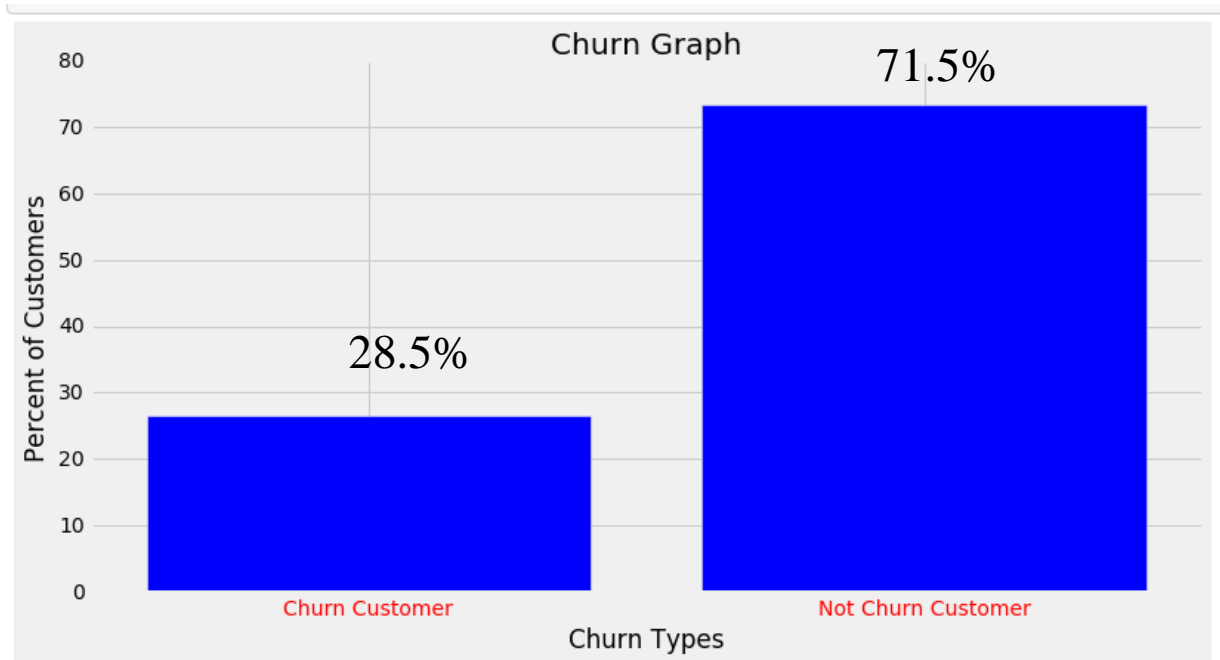
▶ Convert columns with yes/ no to 1/0

▶ When a column has multiple values, I use get_dummy function 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

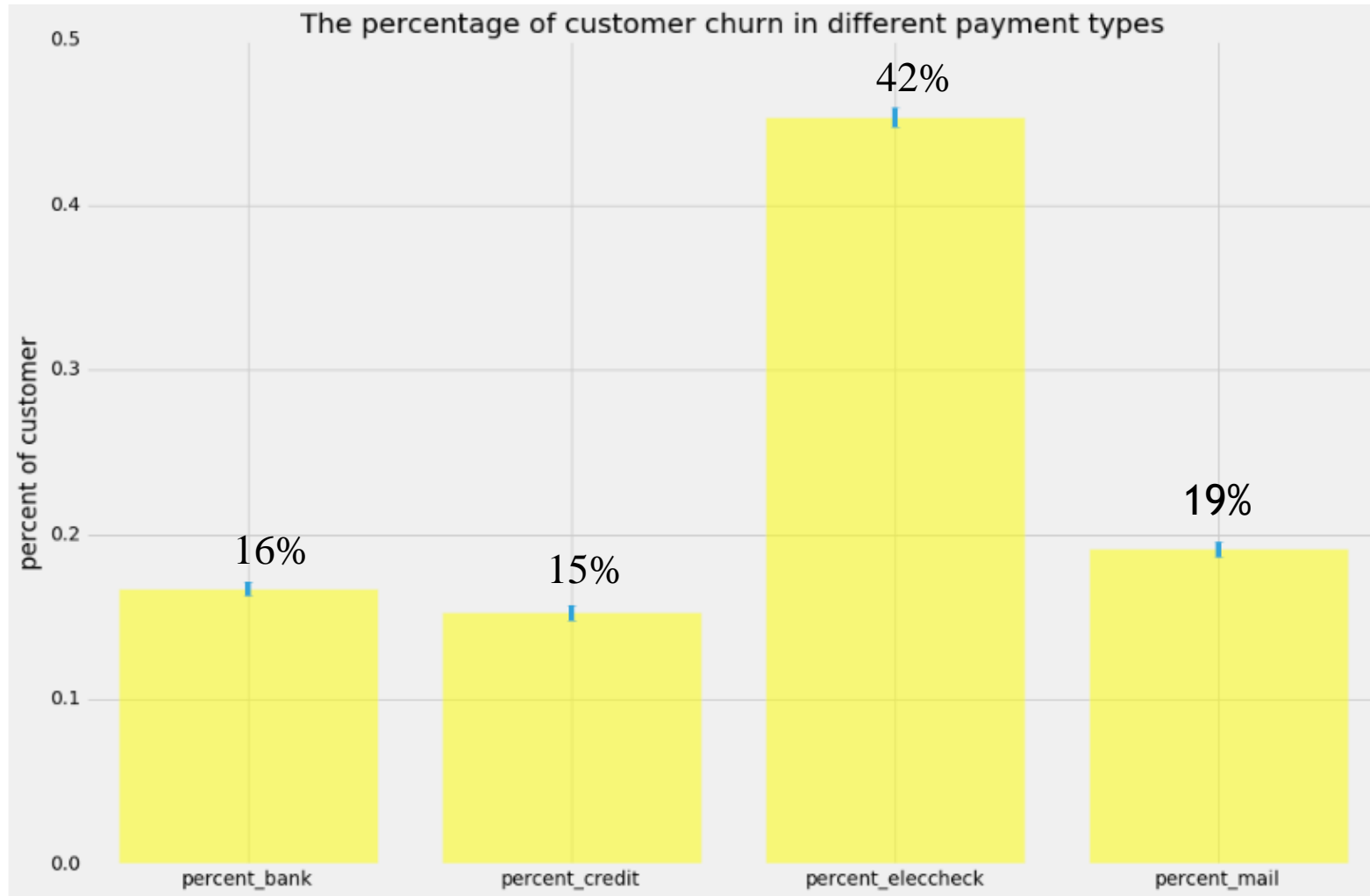# Exploring the Data

Imbalanced classes

Services:

The distribution of phone and internet services through all customers

The percentage of customer churn from using phone and internet services:
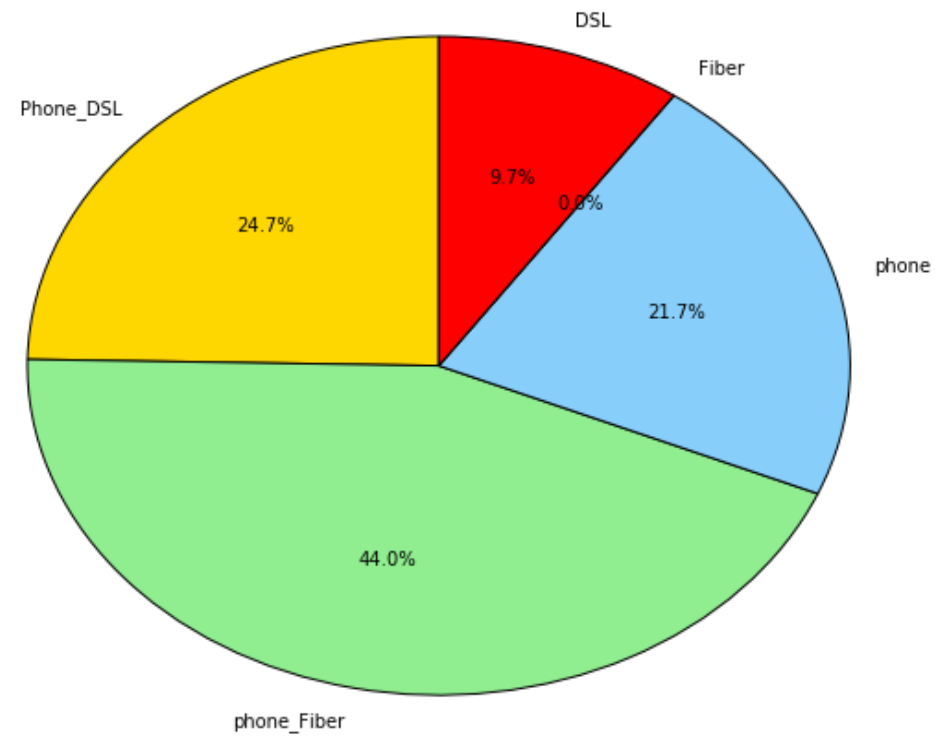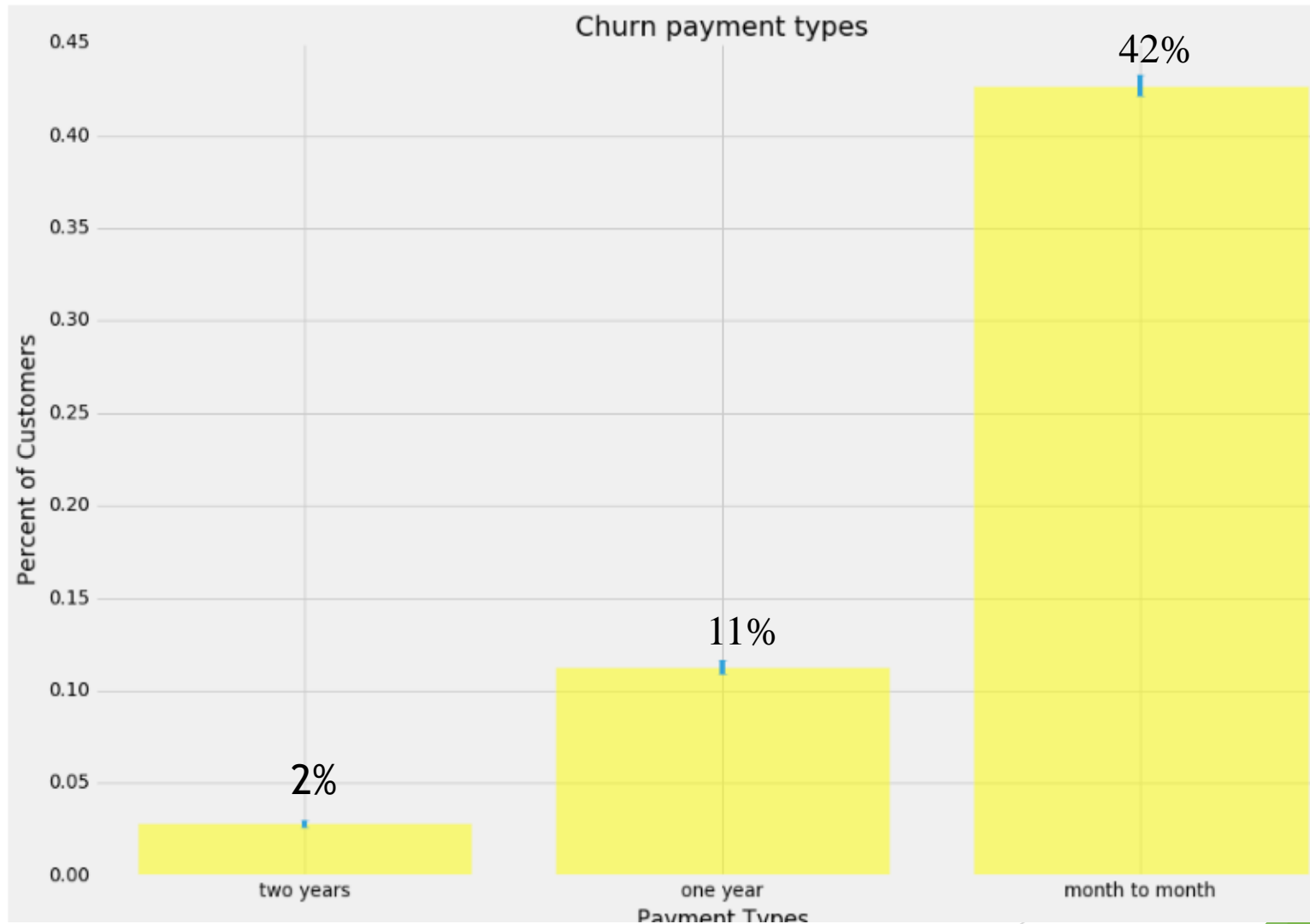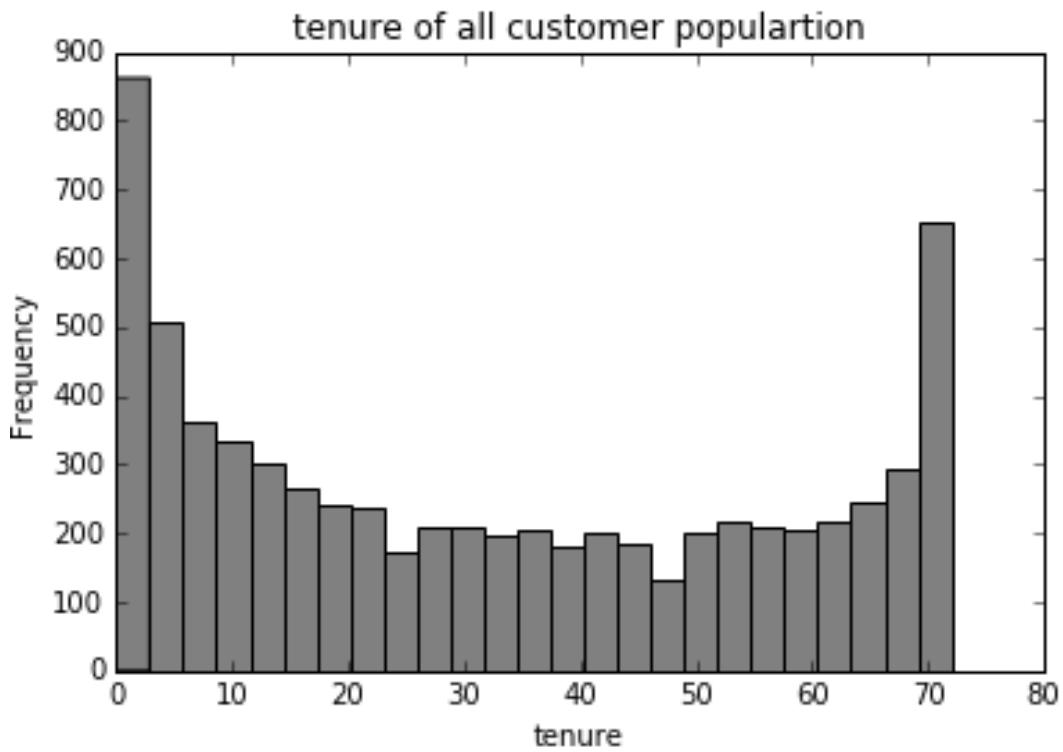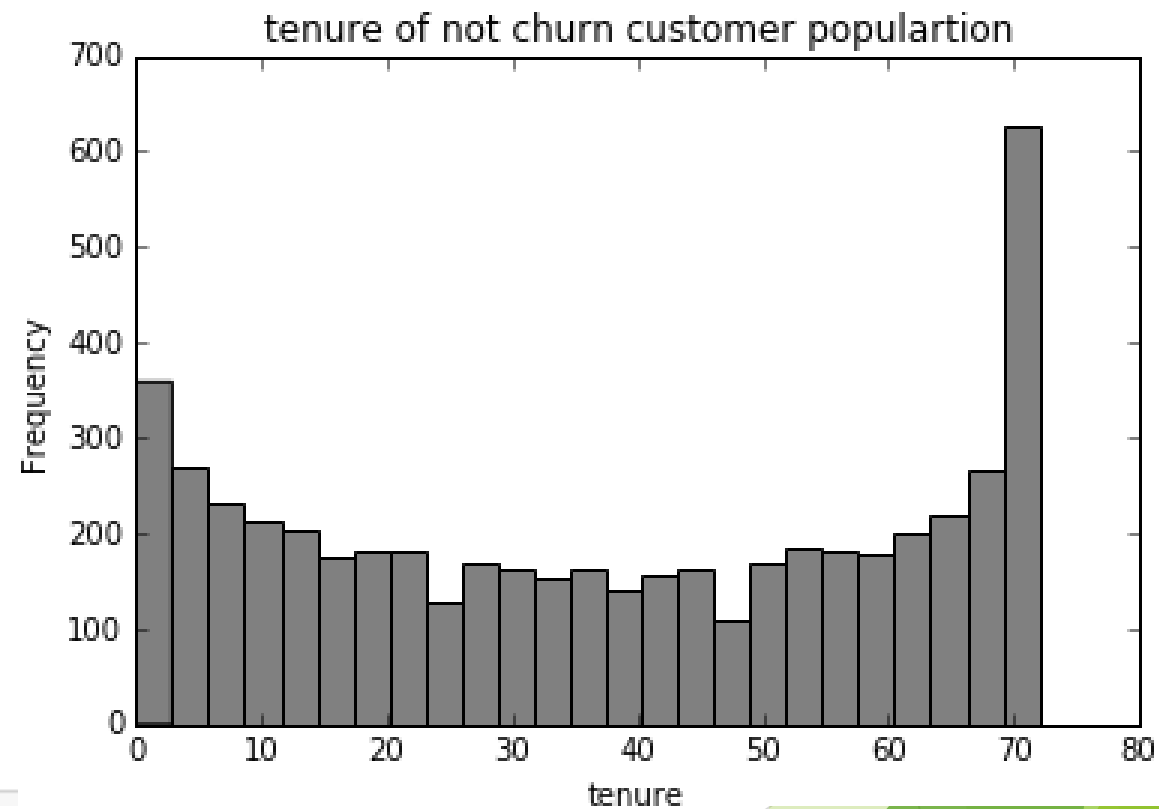

The percentage of customer churn from using phone and internet services

Payment types:
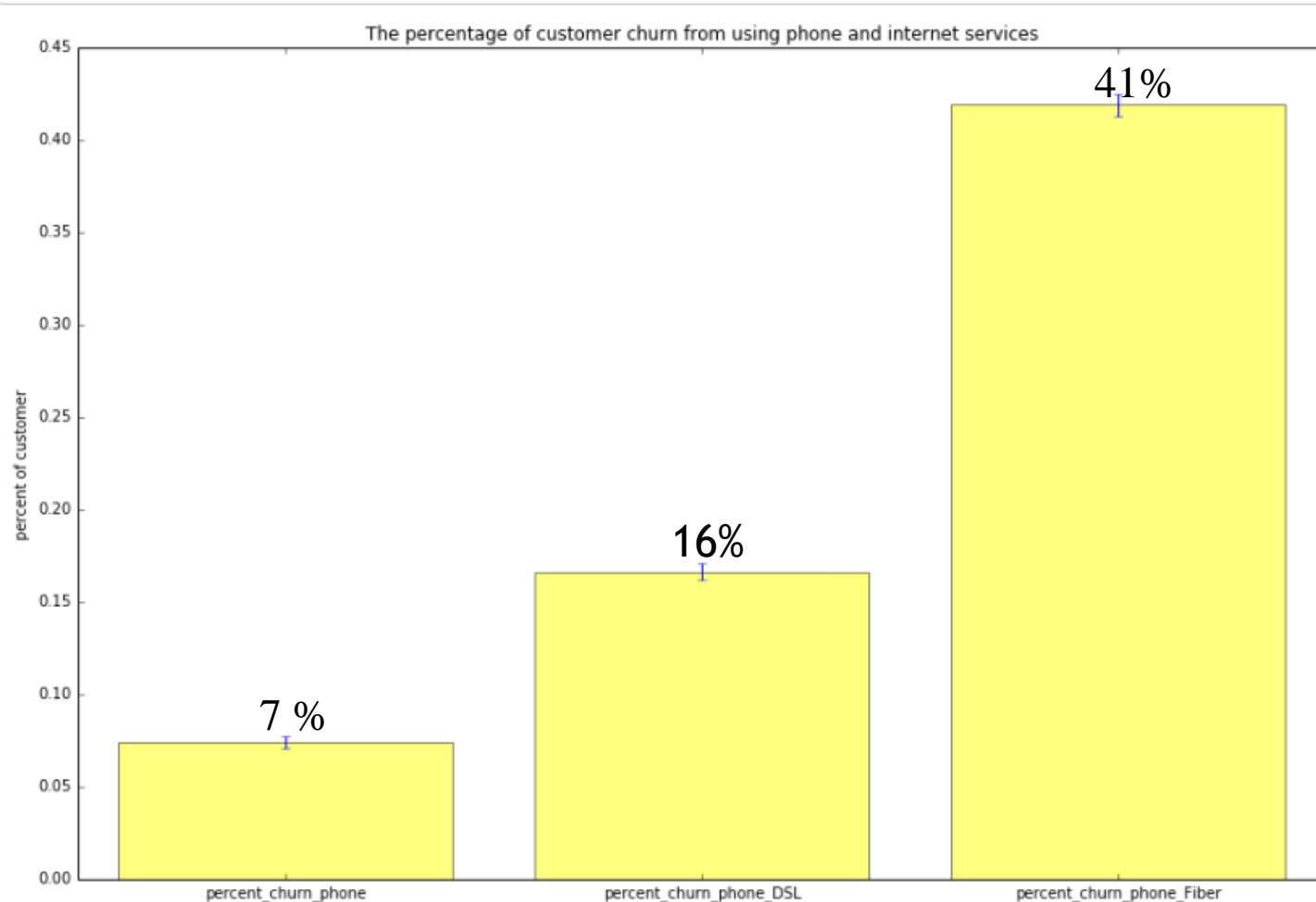
The percentage of customer churn in different payment types

Payment types:

Tenure:

tenure of all customer populartion

Tenure:

tenure of churn customer populartion

tenure of not churn customer populartion
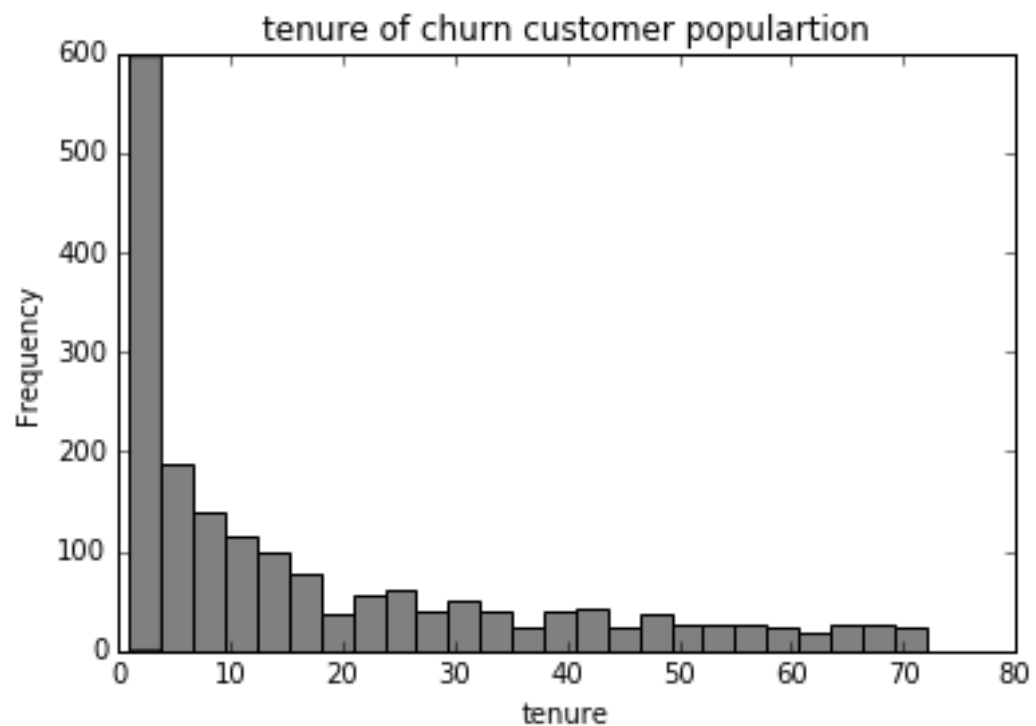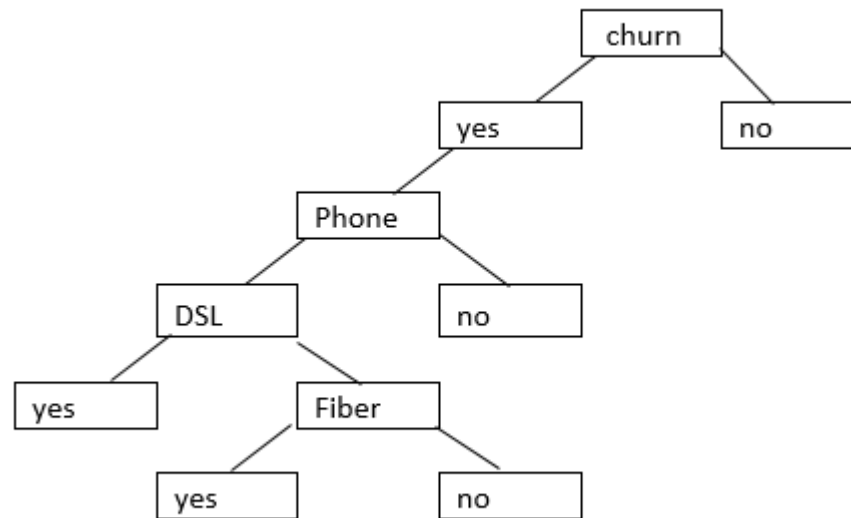
Machine Learning:

- Discover Data Trends

- Predict the best way to increase profit

- Evaluate data analysis

- Marketing purpose

Machine learning:

- The data set is an imbalanced problem.

- We measure model performance through different errors and concepts; such as Precision, recall.

- 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) **(PPV).**

- Recall is also sort of like accuracy but it looks only at the data that is "relevant" in some way. **(TPR)**
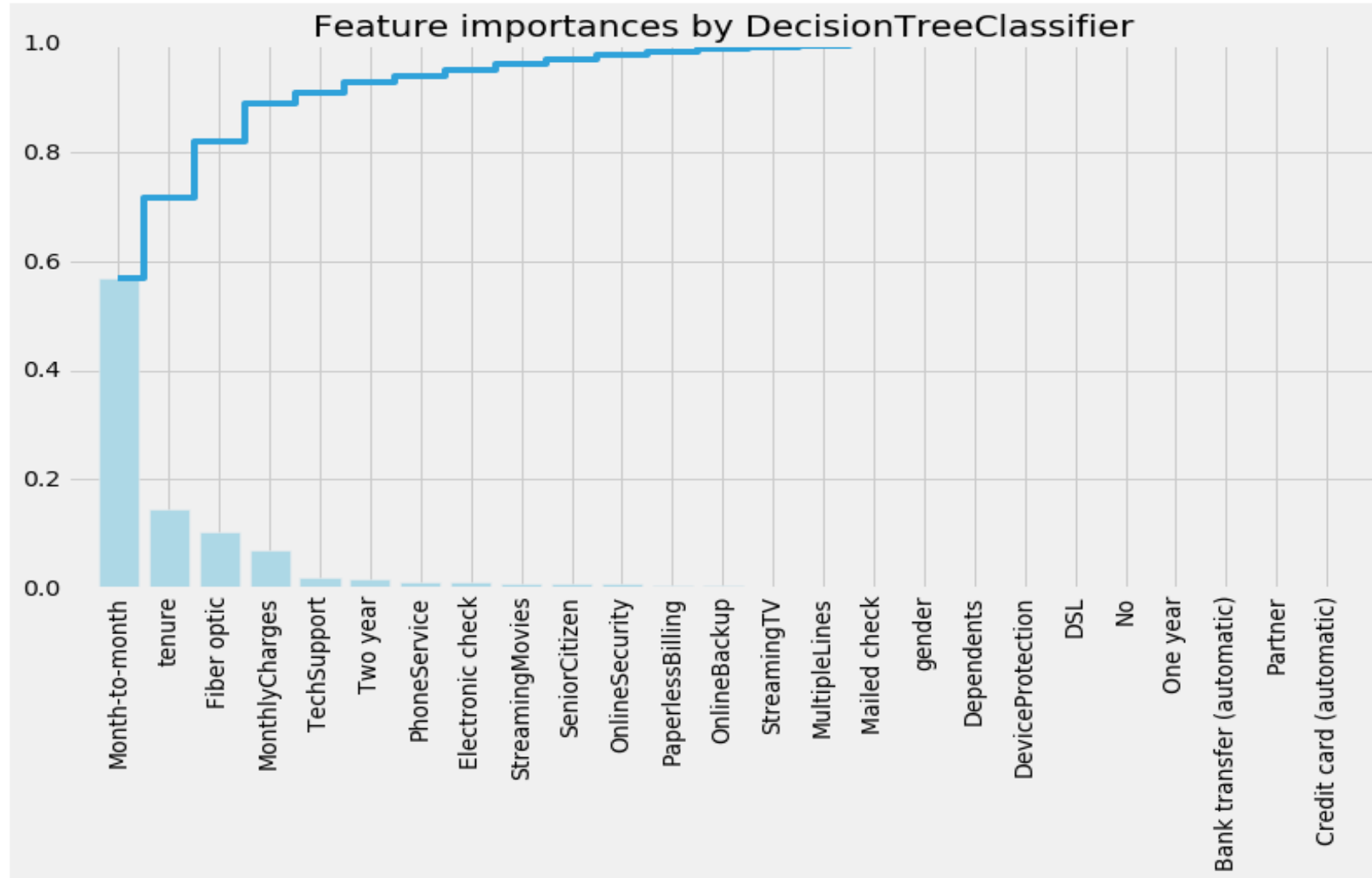
Decision tree:

Random forest:

Random Forest Simplified

Feature important:

Feature importances by DecisionTreeClassifier
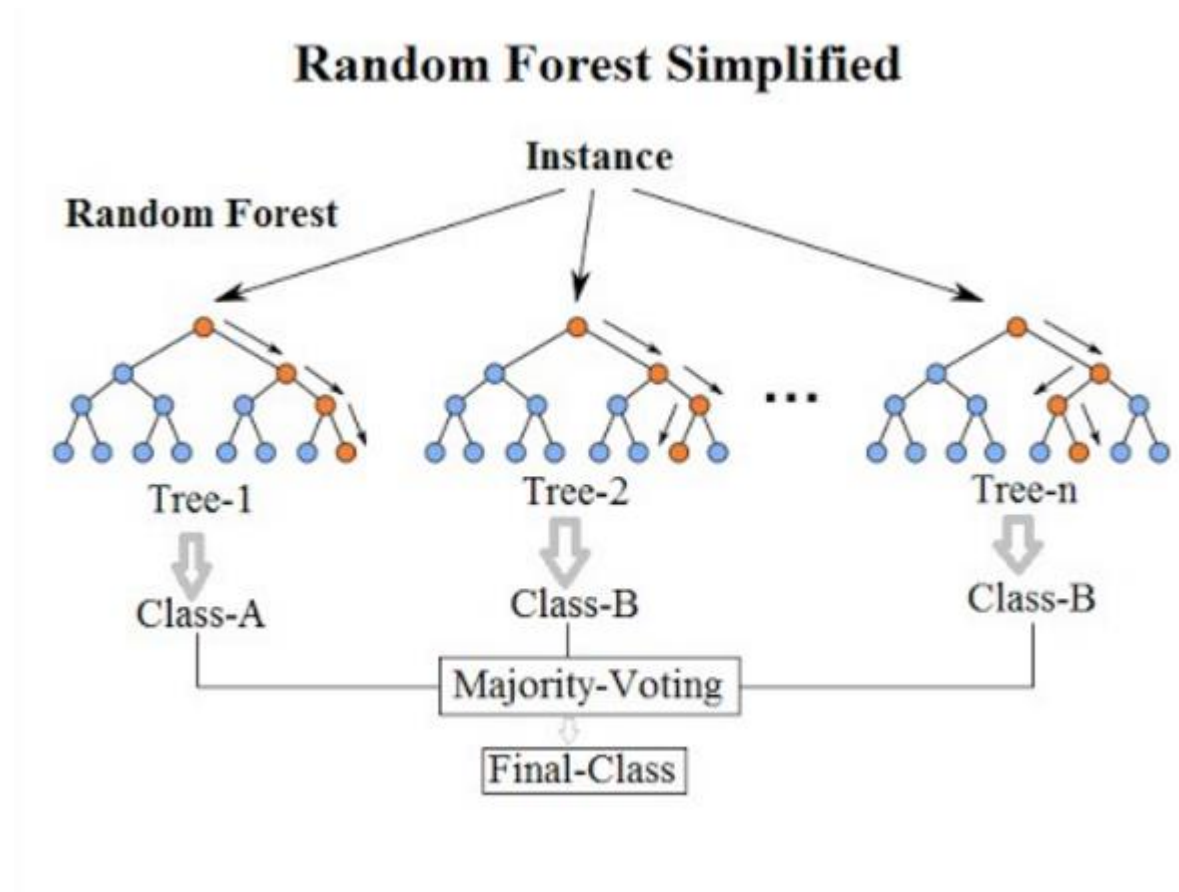
Logistic regression:

Linear Regression:

- Simple:

$$y = b_0 + b_1 * x$$

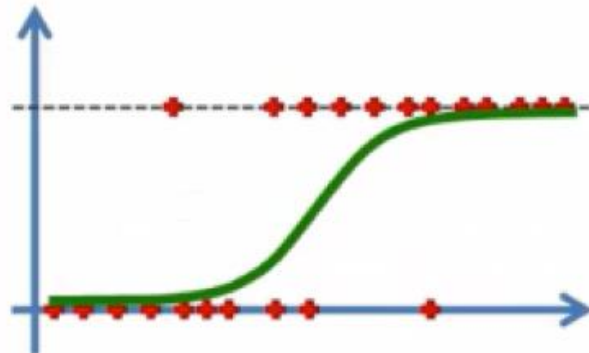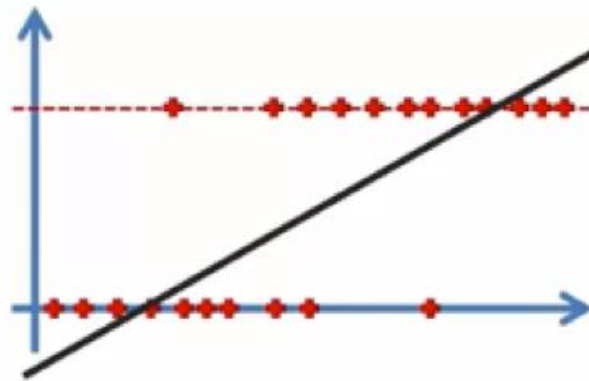**- Multiple:**

$$y = b_0 + b_1 * x_1 + \ldots + b_n * x_n$$
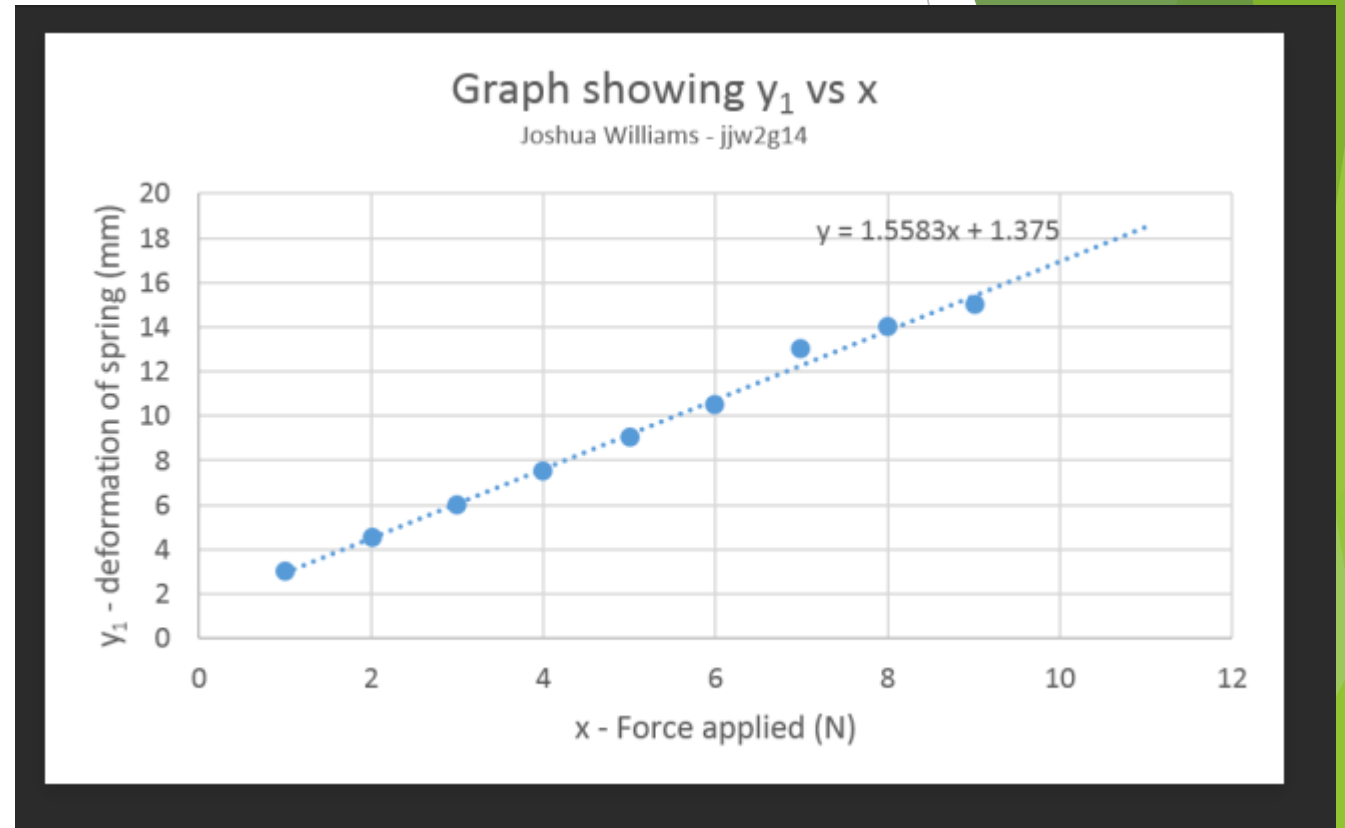
Graph showing y_1 vs x

Joshua Williams - jjw2g14

$y = 1.5583x + 1.375$

y_1 - deformation of spring (mm)

x - Force applied (N)

Logistic regression:
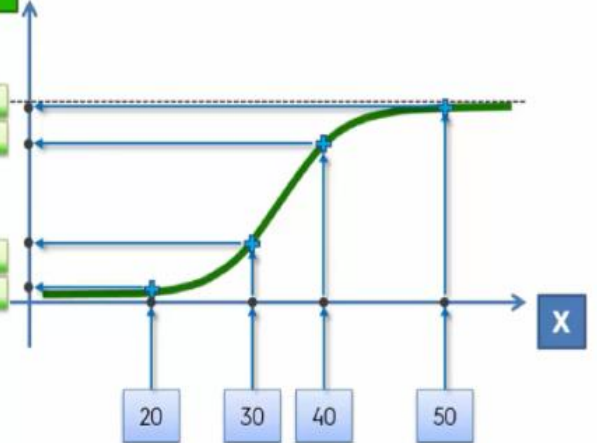
Probability of customer churn:

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

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

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)
```
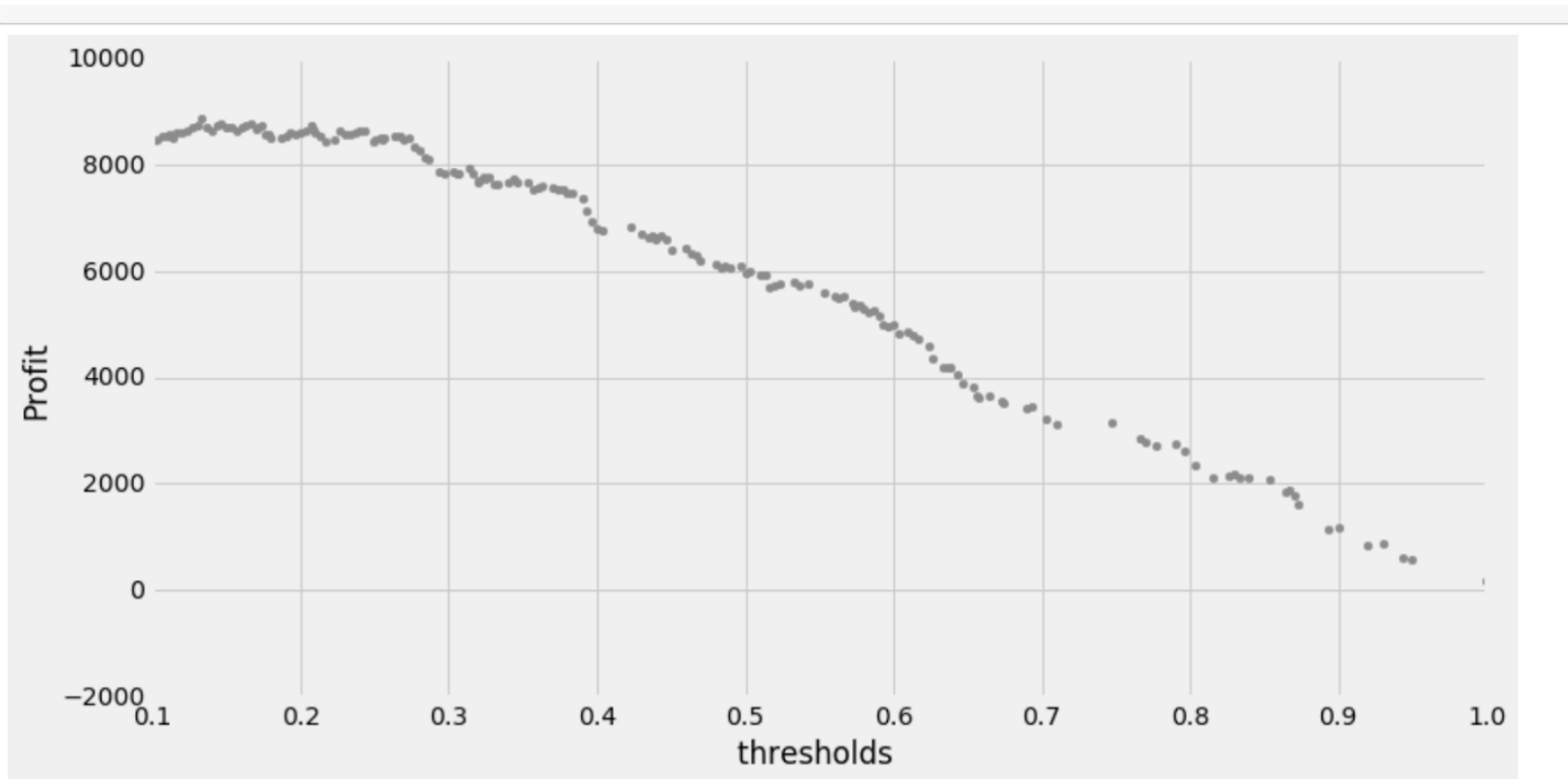
0.801285477303

Probability of customer churn:

Customer ID	Fiber	Tenure	month to Month	probability
7590-VHVEG	0	1	1	-0.6494
5575-GNVDE	0	34	0	-2.8668
3668-QPYBK	0	2	1	-0.6802
7795-CFOCW	0	45	0	-3.2056
9237-HQITU	1	2	1	0.8321

Profit curve

Labor is $20
The customer lifetime value is $100

ROC curve:

ROC Graph

Logistic Regression (area = 0.77)
Random Forest (area = 0.74)
Decision Tree (area = 0.76)
AdaBoost (area = 0.71)
Base Ratek--
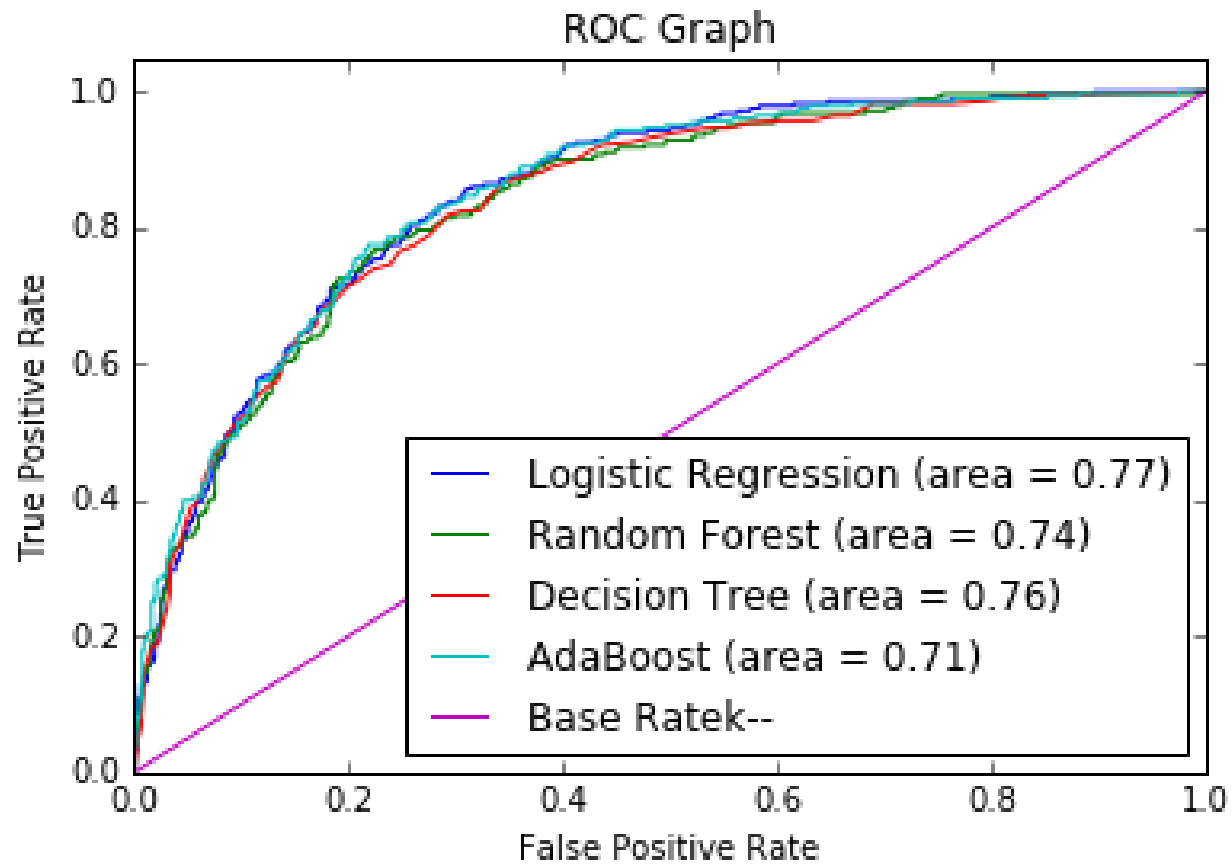
Conclusion:

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.month to month, tenure, and fiber optic are the three most significant features in determining churn.