Logistic Regression

Logistic Regression

- In both the tables, Age and Gender are independent variable.
- In Table 1, we are trying to predict the revenue which is a continuous dependent variable.
- In Table 2, we are trying to predict the probability of subscription based on gender and age which is a binary dependent variable.

Age	Gender	Revenue
65	MALE	69806
23	MALE	25256
53	FEMALE	14091
45	MALE	17176
49	FEMALE	45134
38	FEMALE	38106
32	FEMALE	30865
22	FEMALE	31838
39	MALE	37286
18	FEMALE	36391

Predict the revenue amount using linear regression.

Age	Gender	Subscription
65	MALE	1
23	MALE	1
53	FEMALE	0
45	MALE	0
49	FEMALE	1
38	FEMALE	0
32	FEMALE	1
22	FEMALE	1
39	MALE	0
18	FEMALE	0

Predict the probability of subscription using Logistic regression.

Logistic Regression

- **Logistic regression** is a statistical method for analysing a dataset in which there are one or more independent variables that determine a binary outcome.
- Logistic Regression is a classification algorithm.
- It predicts the probability of occurrence of an event by fitting data to a logit function.
- In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, Success, Pregnant, etc.) or 0 (FALSE, Failure, Non-Pregnant, etc.)
- We use dummy variables to represent binary / categorical outcome, for example we will use “1” for success in exam and “0” for failure in exam.

Continuous Vs. Categorical Variable

- General linear regression model
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
- Independent variable(x's):
- Continuous: Age, income, height -> Uses Numerical value
- Categorical: gender, city, ethnicity -> Uses dummies for example: For Male use "0" and for female "1"

- Dependent Variable (y):
- Continuous: consumption, time spent -> Uses Numerical value
- Categorical: Yes/No -> Uses dummies

Representing the binary Outcome

- The binary outcome have two potential values, usually an observation belongs to a certain category or has satisfied some particular attribute going to be a yes/no question.
- Create a dummy variable to indicate an observation is Yes or No:

  If the answer is Yes dependent variable is 1

  If the answer is No dependent variable is 0
- If we code our dependent variable the other way around (Yes is “0” and No is “1”) the coefficients are going to have the same magnitudes but the opposite signs.

For example:

- In case of student admission, we are interested in finding out the probability of a student obtaining a place on a post-graduate course given the marks from undergraduate degree and the institution they attended.
- In the context of politics we could assess if this person will vote against or in favour of a particular law.
- In a retail context, a customer buy or did not buy a product.
- An individual transactions, use a binary outcome to model if that particular transaction is legal or fraudulent.

Example

- Netflix conducted a marketing activity on its 500 customers out of which some customers subscribed the channel whereas some did not. Now, Netflix wants to analyse the success of their marketing campaign. They have taken a sample of 20 customers and want to analyse the results.
- Subscribe: Indicates a customer has subscribed to a magazine.
- Age(Continuous variable): Examine how age influences the likelihood of subscription

Age	Subscription
62	1
18	0
40	0
51	1
37	1
47	1
32	0
49	1
55	1
52	1
52	1
33	1
41	0
44	0
51	1
52	1
36	0
35	0
30	0
39	0

A linear Model?

- For the above model we can also use the linear model. Only problem we may face is that the dependent variable is binary instead of continuous.
- If we want to use the linear model for this problem , then we need to change the variable “No” to “0” and variable “Yes” to “1” and whenever customer changing from 0 to 1, it increases the likelihood of subscription.

- So, then we can run a simple linear model

$$Subscribe = \beta_0 + \beta_1 * age + \varepsilon$$

Result of Linear Model

- We solved this model using Linear Regression function using Data Analysis Tool pack in Excel

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-0.87	0.37	-2.37	0.03	-1.63	-0.10
Age	0.03	0.01	3.99	0.00	0.02	0.05

- The estimated model is $\mathbf{Subscribe} = -\mathbf{0.86} + \mathbf{0.03} * \mathbf{age}$

Interpretation of Result

- If our dependent variable is binary, then we want to see what makes it change from 0 to 1.
- This can be interpreted as what increases the likelihood of subscription, or P(subscription = 1), which we can also simply denote as p.
- The result can be interpreted as:

$$p\ (subscribe = 1) = p = \text{-0.866 + 0.03 * age}$$

- Every additional year of age increases the probability of subscription by 3%.

Problems with the linear Approach

- The Probabilities are bounded between $(0 \leq p \leq 1)$
- The range of age in our data is between $18 \leq age \leq 62$ so, the youngest customer is 18 year old and the oldest customer is 62 year old.
- It only makes sense to develop a forecasts for observations similar to the ones we have in our data
- Lets assume that the probability of a 40 year old person subscribe is:

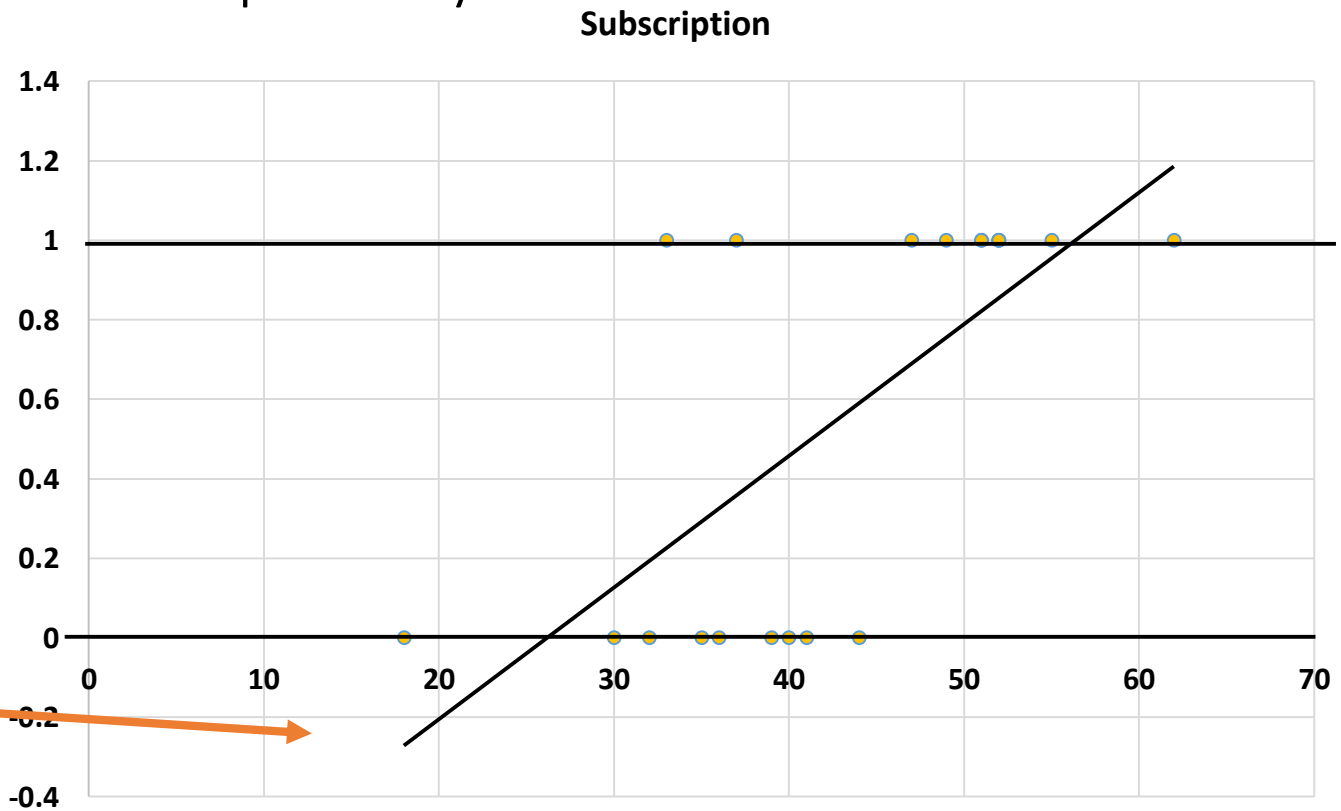
$$\mathrm{P} = -0.866 + 0.03 * 40 = 0.334$$

- What about people with 26 and 57 years of age?

If we plug in 26 we find that the probability that this customer buys is estimated to be -0.005 and this cannot be correct since a probability **cannot have a negative value.**	$\mathrm{P} = -0.866 + 0.03 * 26 = -0.005$
If we plug in 57 we end up with the number of 1.01 which is greater than 1 on came **an invalid value for probability this becomes more clear**.	$\mathrm{P} = -0.866 + 0.03 * 57 = 1.01$

Linear Model

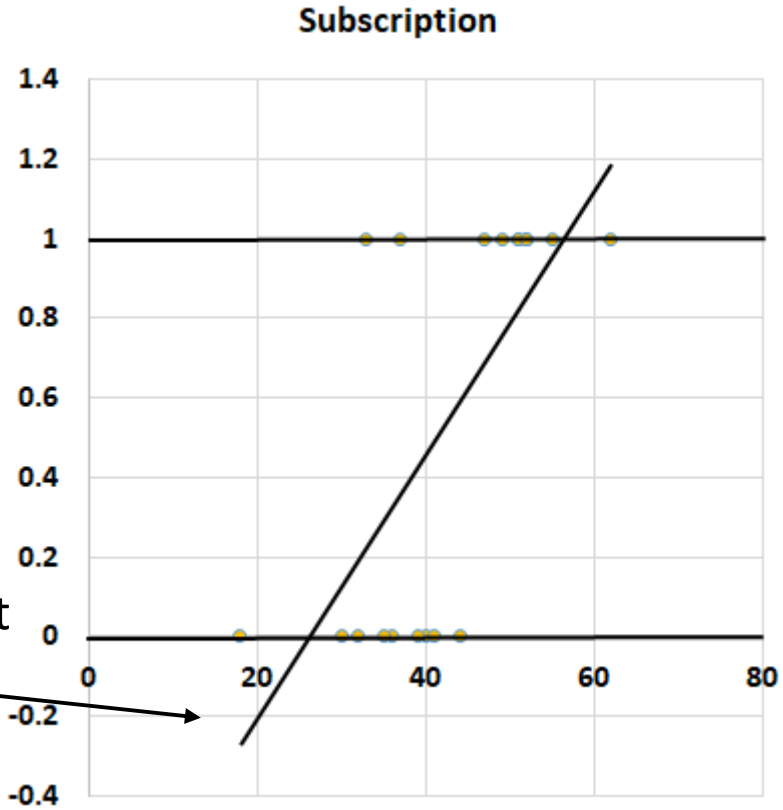
- If we plot the observation, the probabilities should go from 0 to 1 but considering the Netflix example, lets say If the customers are young, below 27 years of age the estimated probabilities are observed to be negative.
- Meanwhile if the customer has more than 57 years of age the estimated probabilities are greater than 1.
- The below model is not working, how could we fix this one opportunity to artificially cap the linear model and say whenever the estimator probability below 0 make it 0 and whenever the estimated probability is above 1 make it 1.



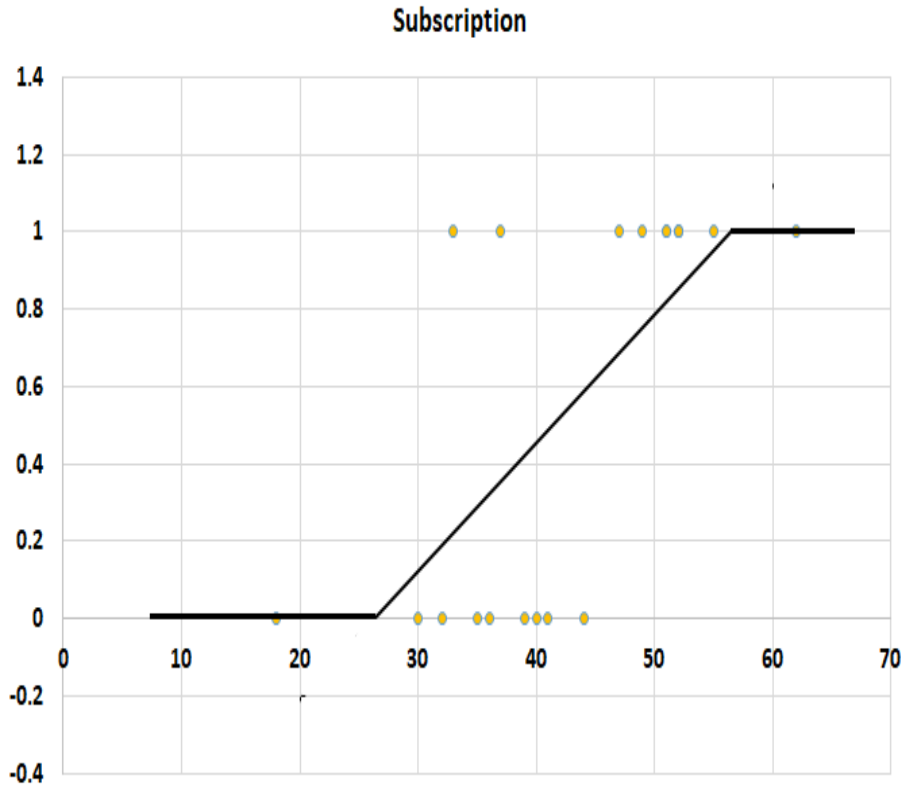
The intercept is -0.08

Linear Model

- The one shown with those breaks in the function but this is two engineered way to custom to be a standard approach
- Could we do something better and let's think what should we do to fix this again note that probabilities should be between 0 and 1



The intercept is -0.08



Fixing the Prior Approach

- We need to somehow constrain p such that $0 \leq p \leq 1$
- We know p = f(age), but the linear function didn't work.
- What must f(.) satisfy to always produce reasonable forecasts?
- f(.) must satisfy two things:
  - ✓It must always be positive (since $p \geq 0$)
  - ✓It must be less than 1 (since $p \leq 1$)

Two Steps!

- Lets try to develop a new function that will satisfy these two criteria
- **It must always be positive (since p $\geq$ 0)**
- What functions could give you a positive numbers
  - ✓The absolute value of a number
  - ✓The squared version of number
- The alternative to this is an exponential form
- $p = \exp(\beta_0 + \beta_1 * age) = e^{(\beta_0 + \beta_1 * age)}$
- For example if $(\beta_0 + \beta_1 * age)$is -2, then exp(-2) = 0.136 (Use excel function "exp" to find exponential value.
- **It must be less than 1 (since p $\leq$ 1)**
- $p = \frac{\exp(\beta_0 + \beta_1 * age)}{\exp(\beta_0 + \beta_1 * age) + 1} = \frac{e^{(\beta_0 + \beta_1 * age)}}{e^{(\beta_0 + \beta_1 * age)} + 1}$

- For example if $\exp(\beta_0 + \beta_1 * age)$ is 1.2 , to make it less than one , we can do : 1.2/(1.2+1) = 1.2/2.2

The linear thinking is not completely gone

- The previous expression (by doing some algebra) can be rewritten as:
- $\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * \text{age}$
- P being the result of the prior expression is equal to a linear function of age that looks just like the linear simple regression models.
- Even though the probability of the customer subscribing (p) is not linear function of age, we can perform a simple transformation on it such that it is now a linear function of age.
- The above equation is used in **Logistic Regression.**

Coefficients and Standard Error

- Here, the dependent Variable is Subscribe
- The sign of coefficient represents the positive or negative influence on the dependent variable.
- The coefficient of age is 0.242 which is positive
- The positive sign of age coefficient represents that as the age grows customers are more likely to subscribe the Netflix.

Standard Error

- Standard Error is used to find out the confidence interval of the model

  Coefficient Value $\pm$ 2* Standard Error
- Here, the confidence interval is (0.242-2*0.1, 0.242+2*0.1) = (0.042,0.442)

	coeff b	s.e.	Wald	p-value	exp(b)	lower	upper
Intercept	-10.021	4.210	5.665	0.017	0.000		
Age	0.242	0.100	5.877	0.015	1.273	1.047	1.548

The estimated Logistic Model

- It can be see that the coefficient for the intercept and the slope is -10.021 and 0.242 but how do we interpret these coefficients are different
- The estimated model was
- $\ln\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right) = -10.021 + 0.242 * \mathrm{age}$
- For every unit increase of age $\ln\left(\frac{\mathrm{p}}{1-\mathrm{p}}\right)$ increases by 0.242 units.
- Or its written in terms of the probability p we have:
- $p = \frac{\exp(-10.021+0.242*age)}{\exp(-10.021+0.242*age)+1} = \frac{e^{(-10.021+0.242*age)}}{e^{(-10.021+0.242*age)}+1}$
- Lets check the probabilities of subscription if the age is 18 year and 33 year.
- Probability for 18 year old is $\mathrm{p} = \frac{\mathrm{e}^{(-10.021+0.242*18)}}{\mathrm{e}^{(-10.021+0.242*18)}+1} = 0.0035$
- Probability for 33 year old is $\mathrm{p} = \frac{\mathrm{e}^{(-10.021+0.242*33)}}{\mathrm{e}^{(-10.021+0.242*33)}+1} = 0.1156$

Logistic Regression Output

- In the estimated model, the dependent variable is the $\ln(\frac{p}{1-p})$ is -10.021 + 0.242* age.
- **Success** - Number of times the observed subscription is 1 for a particular age.
- **Failure** - Number of times the observed subscription is 0 for a particular age.
- **Total** - is the sum of Success and Failure.
- **P-obs** - For a age the subscription is 1.
- **P-Pred** - The predicted probability of the p.
- **Suc-Pred** – Suc-Pred is same as p-Pred. It is the successful predicted probability.
- **Fail-Pred** – It is 1minus Suc-Pred.
- **%Correct** – It is another way to gauge the fit of the model to the observed data. The statistic tells that 85% of the observed cases are predicted accurately by the model.

Logistic Regression								
Age	Success	Failure	Total	p-Obs	p-Pred	Suc-Pred	Fail-Pred	% Correct
18	0	1	1	0	0.003	0.003	0.997	100
30	0	1	1	0	0.059	0.059	0.941	100
32	0	1	1	0	0.092	0.092	0.908	100
33	1	0	1	1	0.114	0.114	0.886	0
35	0	1	1	0	0.173	0.173	0.827	100
36	0	1	1	0	0.210	0.210	0.790	100
37	1	0	1	1	0.253	0.253	0.747	0
39	0	1	1	0	0.355	0.355	0.645	100
40	0	1	1	0	0.412	0.412	0.588	100
41	0	1	1	0	0.471	0.471	0.529	100
44	0	1	1	0	0.648	0.648	0.352	0
47	1	0	1	1	0.792	0.792	0.208	100
49	1	0	1	1	0.860	0.860	0.140	100
51	2	0	2	1	0.909	1.818	0.182	100
52	3	0	3	1	0.927	2.781	0.219	100
55	1	0	1	1	0.963	0.963	0.037	100
62	1	0	1	1	0.993	0.993	0.007	100
	11	9	20			11	9	85

- In our example, the cut off value 0.5 i.e. 50% which means that if the success predicted probability is more than 50% then it will considered as very high chances of subscription. And if it is less than 50% then it will considered as very less chances of subscription.
- As shown in Table, the success predicted probability for 18 year old person is 0.003 i.e. 0.3% means that the person is not going to subscribe for Netflix is TRUE. Hence, the failure should be 1, Hence % correct is 100%.
- As shown in Table, the success predicted probability for 33 yr old person is 0.114 i.e. 11.4% means that the person is not going to subscribe for Netflix is FALSE. Hence, Failure should be 1, but in the example we can see that success is 1 instead of failure. Hence the % correct is 0% means the predicted value of model is wrong.

Logistic Regression								
Age	Success	Failure	Total	p-Obs	p-Pred	Suc-Pred	Fail-Pred	% Correct
18	0	1	1	0	0.003	0.003	0.997	100
30	0	1	1	0	0.059	0.059	0.941	100
32	0	1	1	0	0.092	0.092	0.908	100
33	1	0	1	1	0.114	0.114	0.886	0
35	0	1	1	0	0.173	0.173	0.827	100
36	0	1	1	0	0.210	0.210	0.790	100
37	1	0	1	1	0.253	0.253	0.747	0
39	0	1	1	0	0.355	0.355	0.645	100
40	0	1	1	0	0.412	0.412	0.588	100
41	0	1	1	0	0.471	0.471	0.529	100
44	0	1	1	0	0.648	0.648	0.352	0
47	1	0	1	1	0.792	0.792	0.208	100
49	1	0	1	1	0.860	0.860	0.140	100
51	2	0	2	1	0.909	1.818	0.182	100
52	3	0	3	1	0.927	2.781	0.219	100
55	1	0	1	1	0.963	0.963	0.037	100
62	1	0	1	1	0.993	0.993	0.007	100
	11	9	20			11	9	85

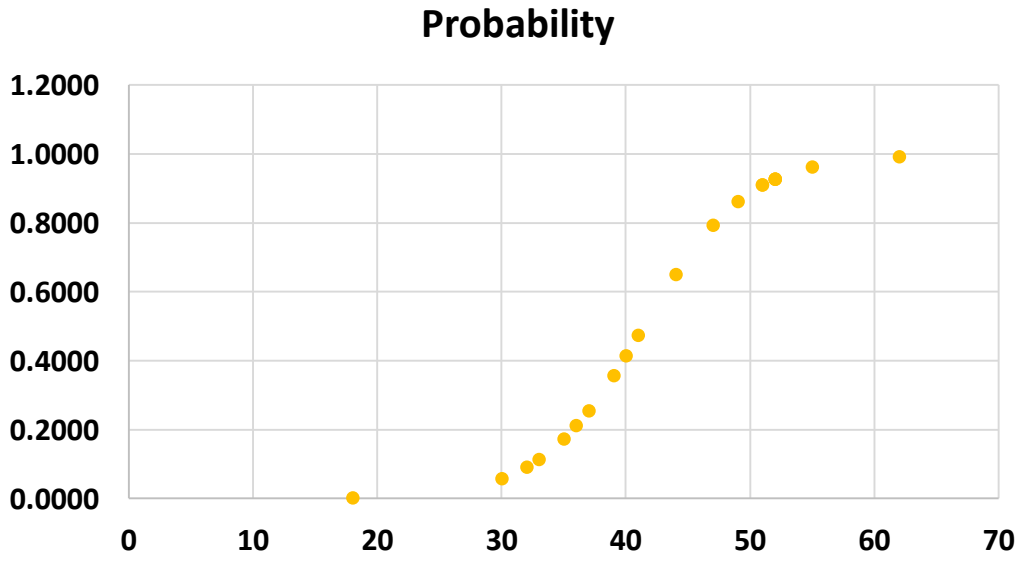
Logistic Model

- Lets See the Probability

Age	Equation = -10.021+0.242*age	exp (Equation)	exp (Equation) + 1	Probability = exp(Equation)/exp(Equation)+1
62	4.983	145.911	146.911	0.9932
18	-5.665	0.003	1.003	0.0035
40	-0.341	0.711	1.711	0.4156
51	2.321	10.186	11.186	0.9106
37	-1.067	0.344	1.344	0.2560
47	1.353	3.869	4.869	0.7946
32	-2.277	0.103	1.103	0.0930
49	1.837	6.278	7.278	0.8626
55	3.289	26.816	27.816	0.9640
52	2.563	12.975	13.975	0.9284
52	2.563	12.975	13.975	0.9284
33	-2.035	0.131	1.131	0.1156
41	-0.099	0.906	1.906	0.4753
44	0.627	1.872	2.872	0.6518
51	2.321	10.186	11.186	0.9106
52	2.563	12.975	13.975	0.9284
36	-1.309	0.270	1.270	0.2127
35	-1.551	0.212	1.212	0.1749
30	-2.761	0.063	1.063	0.0595
39	-0.583	0.558	1.558	0.3582

Logistic Model Plot

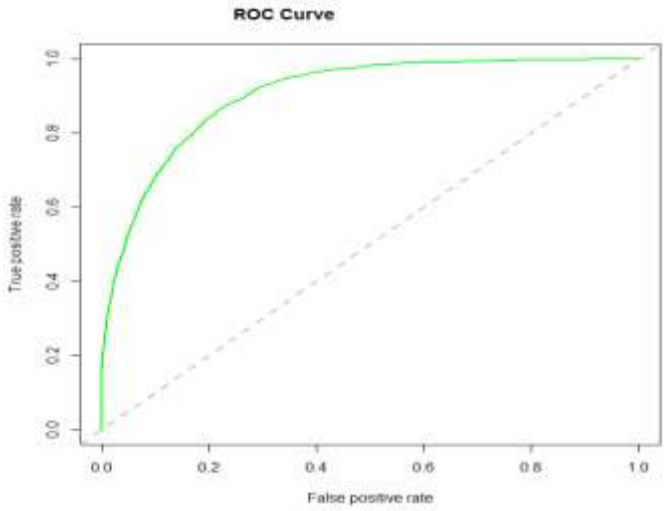
- If we plot the probability against age, we note that the probability is no longer below zero or above one in fact as customers grow older the probability asymptotically gets closer to one
- When customers grow younger the function is asymptotically closer to zero but never below zero or above one
- This is the plot of a logistic model



Model Validation

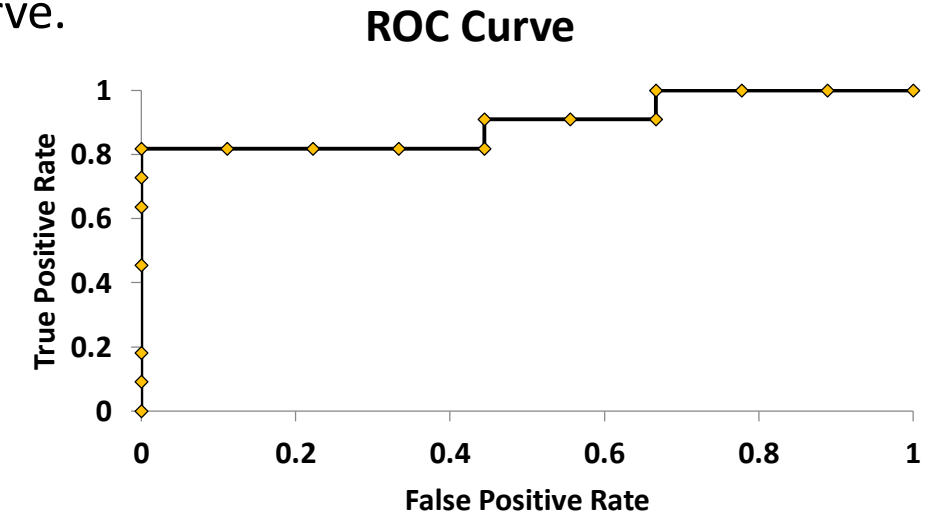
ROC

- Receiver Operating Characteristic(ROC) summarizes the model's performance by evaluating the trade offs between true positive rate (sensitivity) and false positive rate(1- specificity).
- For plotting ROC, assume p > 0.5 since we are more concerned about success rate.
- ROC summarizes the predictive power for all possible values of p > 0.5.
- The area under curve (AUC) is a perfect performance metric for ROC curve it is also referred to as index of accuracy(A) or concordance index.
- Higher the area under curve, better the prediction power of the model.
- The ROC of a perfect predictive model has TP equals 1 and FP equals 0. This curve will touch the top left corner of the graph.
- For model performance, we can also consider likelihood function.
- Likelihood function selects coefficient values which maximize the likelihood of explaining the observed data.
- It indicates goodness of fit as its value approaches one, and a poor fit of the data as its value approaches zero.



ROC

- As mentioned in previous slide, the ROC of perfect predictive model has TP equals 1.
- In our example , the True Positive Rate for ROC curve is approaching to 1.
- Hence the model covering highest area are under the curve.



ROC Table							
p-Pred	Failure	Success	Fail-Cum	Suc-Cum	FPR	TPR	AUC
			0	0	1	1	0.111
0.003	1	0	1	0	0.889	1	0.111
0.059	1	0	2	0	0.778	1	0.111
0.092	1	0	3	0	0.667	1	0
0.114	0	1	3	1	0.667	0.909	0.101
0.173	1	0	4	1	0.556	0.909	0.101
0.210	1	0	5	1	0.444	0.909	0
0.253	0	1	5	2	0.444	0.818	0.091
0.355	1	0	6	2	0.333	0.818	0.091
0.412	1	0	7	2	0.222	0.818	0.091
0.471	1	0	8	2	0.111	0.818	0.091
0.648	1	0	9	2	0	0.818	0
0.792	0	1	9	3	0	0.727	0
0.860	0	1	9	4	0	0.636	0
0.909	0	2	9	6	0	0.455	0
0.927	0	3	9	9	0	0.182	0
0.963	0	1	9	10	0	0.091	0
0.993	0	1	9	11	0	0	0
							0.899

Confusion Matrix

- Confusion matrix is a tabular representation of Actual vs Predicted values.
- It is used to find out the accuracy of the model and avoid overfitting.

- $\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$

		Predicted	
		Good	Bad
Actual	Good	True Positive (TP)	False Negative (FN)
	Bad	False Positive (FP)	True Negative (TN)

- **Sensitivity** or **True Positive Rate** measures the proportion of positives that are correctly identified.
- **Specificity** or **True Negative Rate** measures the proportion of negatives that are correctly identified.
- $\text{True Negative Rate (TNR), Specificity} = \frac{TN}{TN+FP}$
- $\text{False Positive Rate (FPR)}, 1 - \text{Specificity} = \frac{FP}{TN+FP}$

Sum to 1

- $\text{True Positive Rate (TPR), Sensitivity} = \frac{TP}{TP+FN}$
- $\text{False Negative Rate (FNR)}, 1 - \text{Sensitivity} = \frac{FN}{TP+FN}$

Sum to 1

Confusion Matrix

- **The Accuracy is 85% with a cut off of 50%**
- $\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+ TN+FP+FN}} = \frac{9+8}{9+1+2+8} = \frac{17}{20} = 0.85 = 85\%$
- $\text{True Negative Rate (TNR), Specificity} = \frac{\text{TN}}{\text{TN+FP}} = \frac{8}{2+8} = 0.8 = 80\%$
- $\text{False Positive Rate (FPR), 1 – Specificity} = \frac{\text{FP}}{\text{TN+FP}} = \frac{2}{2+8} = 0.2 = 20\%$
- $\text{True Positive Rate (TPR), Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} = \frac{9}{9+1} = 0.9 = 90\%$
- $\text{False Negative Rate (FNR), 1 – Sensitivity} = \frac{\text{FN}}{\text{TP+FN}} = \frac{1}{9+1} = 0.1 = 10\%$
- The calculated accuracy and the model accuracy is same.

Classification Table			
	Suc-Obs	Fail-Obs	
Suc-Pred	9	1	10
Fail-Pred	2	8	10
	11	9	20
Accuracy	0.818	0.889	0.85
Cutoff	0.5		

Thank You.