LOGISTIC REGRESSION

Outline

- About Logistic Regression
- Comparison to Linear Regression
- Logit model and estimation
- Visualization of logit function on binary data
- Prediction and Inference

Logistic Regression

- Logistic regression is a classification technique used when the response variable is categorical.
- It models the log-odds of the probability of an event occurring as a linear combination of input features.

  $$\log(p/(1-p)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

  – *where p refers to p(Y=1|X)*
- The model outputs probabilities and is commonly used for binary classification tasks.

Use Cases

- Predicting customer churn, whether customers will stay or leave
- Modeling likelihood of purchase
- Determine characteristics of online ads that customer click
- Estimating likelihood of loan default
- Disease diagnosis

Strengths

- Provides a probability of the prediction
- Coefficients help in evaluating feature influence
- Easy to interpret and communicate

Weaknesses

- Assumes a linear relationship in the log-odds
- Less effective when there is a class imbalance
- Sensitive to multicollinearity, outliers, and missing data

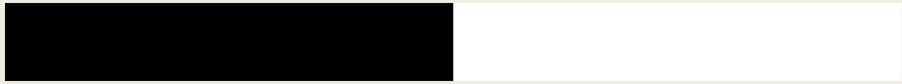
Linear Regression vs. Logistic Regression

Linear Regression

- How much is John Doe going to spend?
- Shades of grey

Logistic Regression

- Will John Doe buy?
- Black or White



Logit Model

- Involves modeling the probability of a binary outcome
- Using conventional linear regression generates meaningless predictions such as probabilities less than 0 or more than 1.
- Therefore, logistic regression uses the logit function to constrain predicted probabilities to be between 0 and 1

$$\log(p/(1-p)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

  - *where p refers to $p(Y=1|X)$*
- Using the logit function ensures $0 < p(Y=1|X) < 1$
- Above equation can be transformed to compute probability

$$p = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p} / (1+e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p})$$

Estimation

- Logit Model is estimated by a technique called Maximum Likelihood.

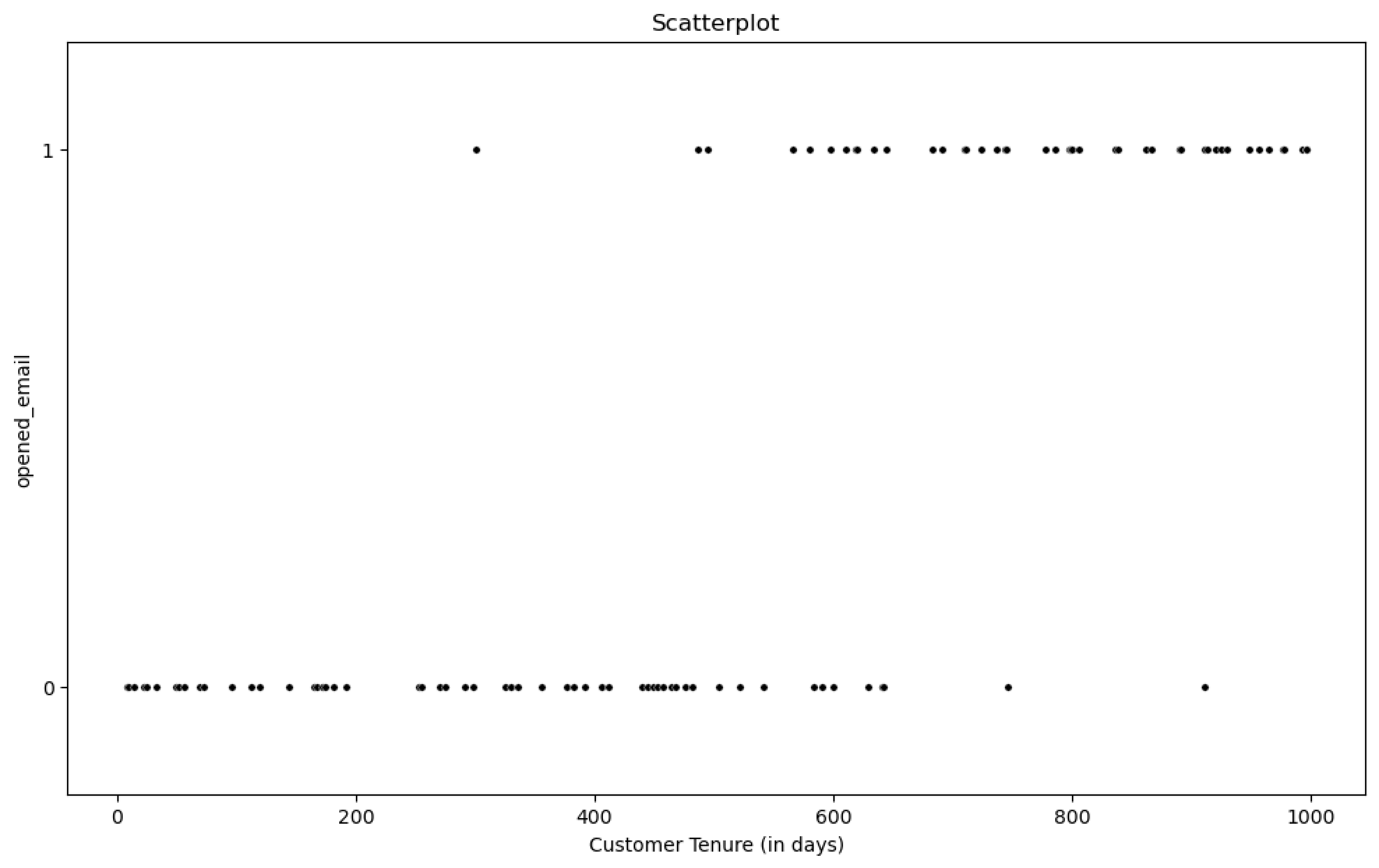
Logit Model Example

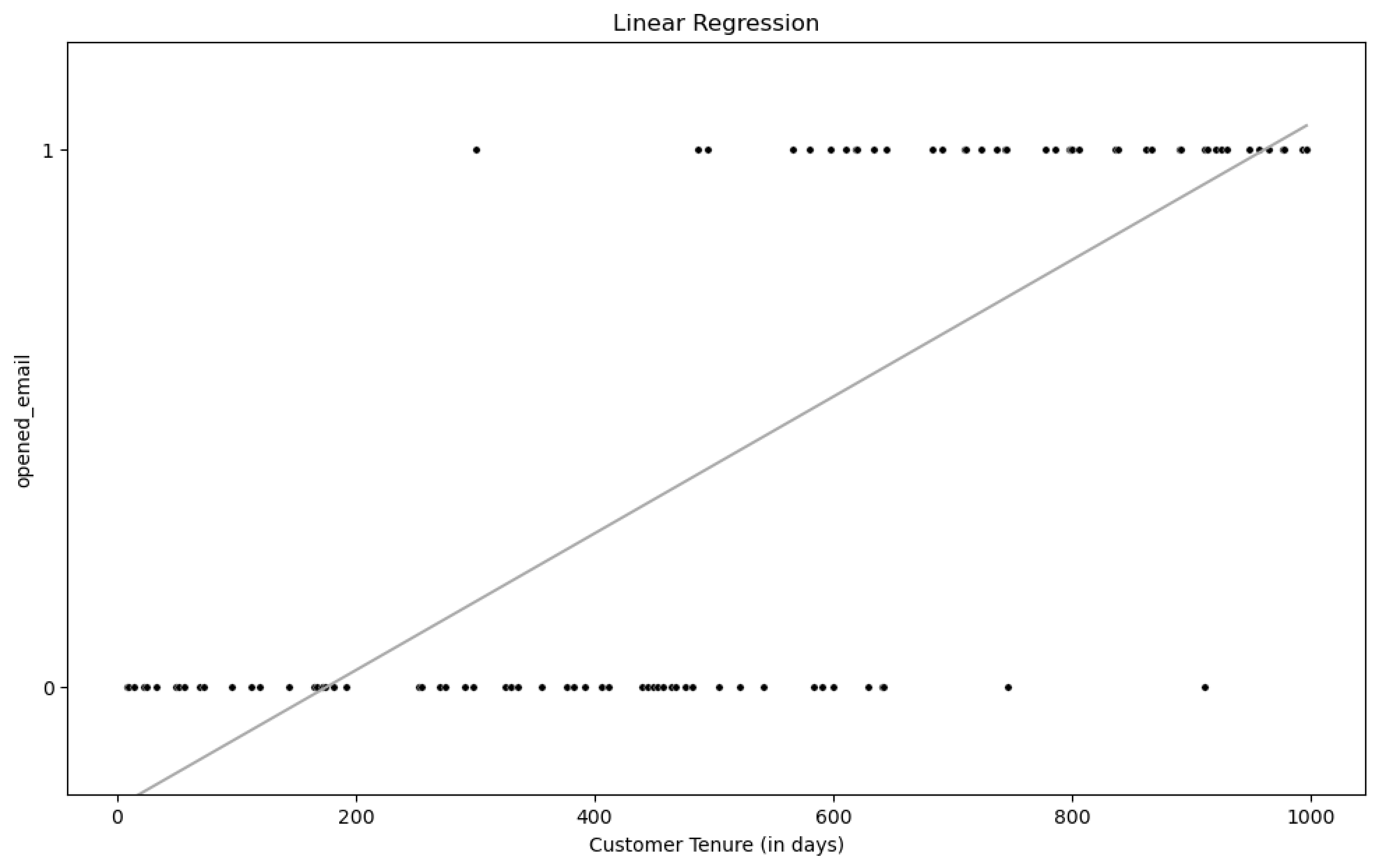
- Imagine a realtor is interested in identifying factors that influence the sale of a house
- Outcome is the Sale of a house in two weeks and p is the probability of sale in two weeks

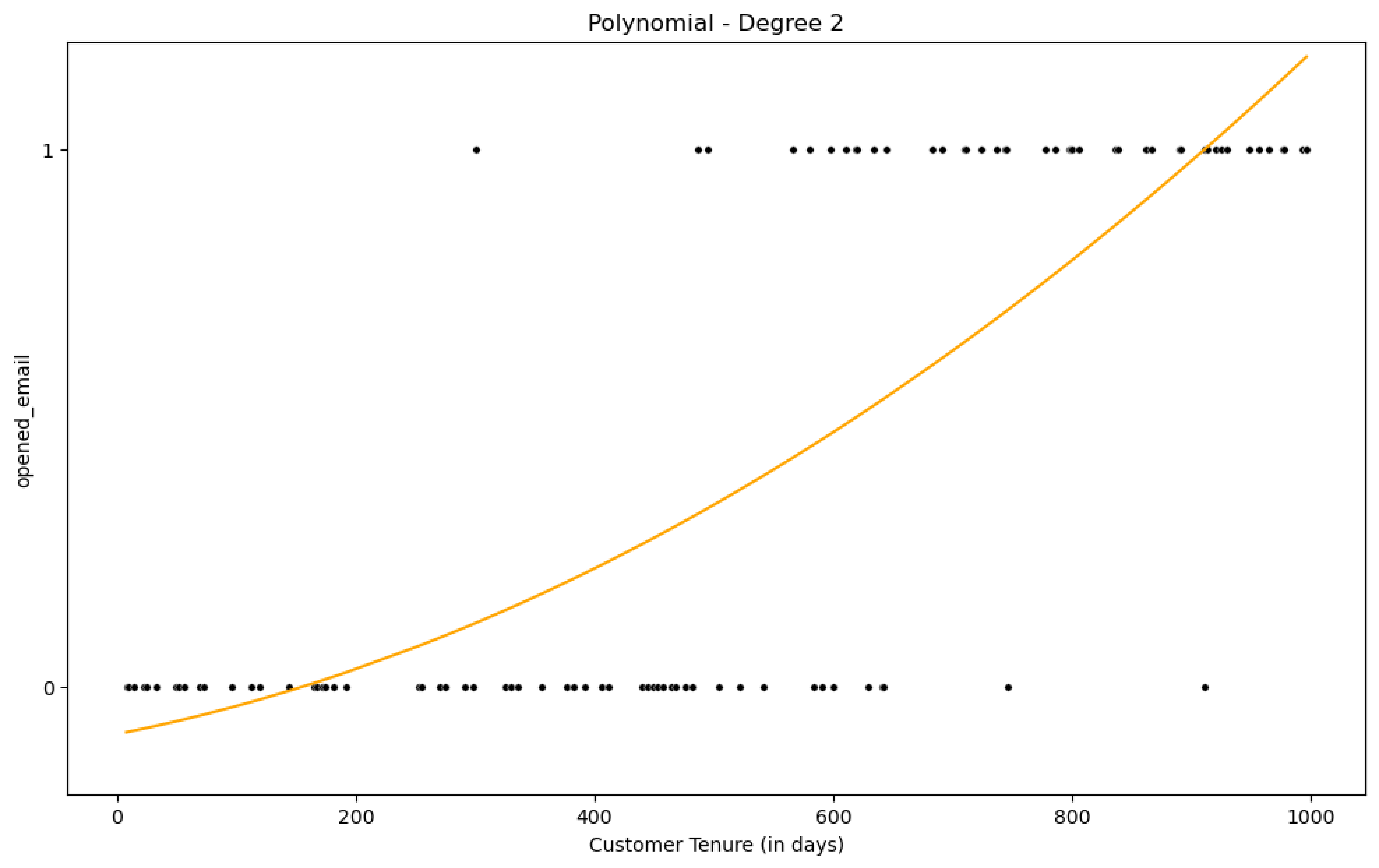
1. Outcome = f(total area, number of rooms, age)
2. ln(p/1-p) = b_0 + b_1*area + b_2*number of rooms + b_3*age
3. p = $e^{\text{b0 + b1*area + b2*number of rooms + b3*age}}$ /1+$e^{\text{b0 + b1*area + b2*number of rooms + b3*age}}$

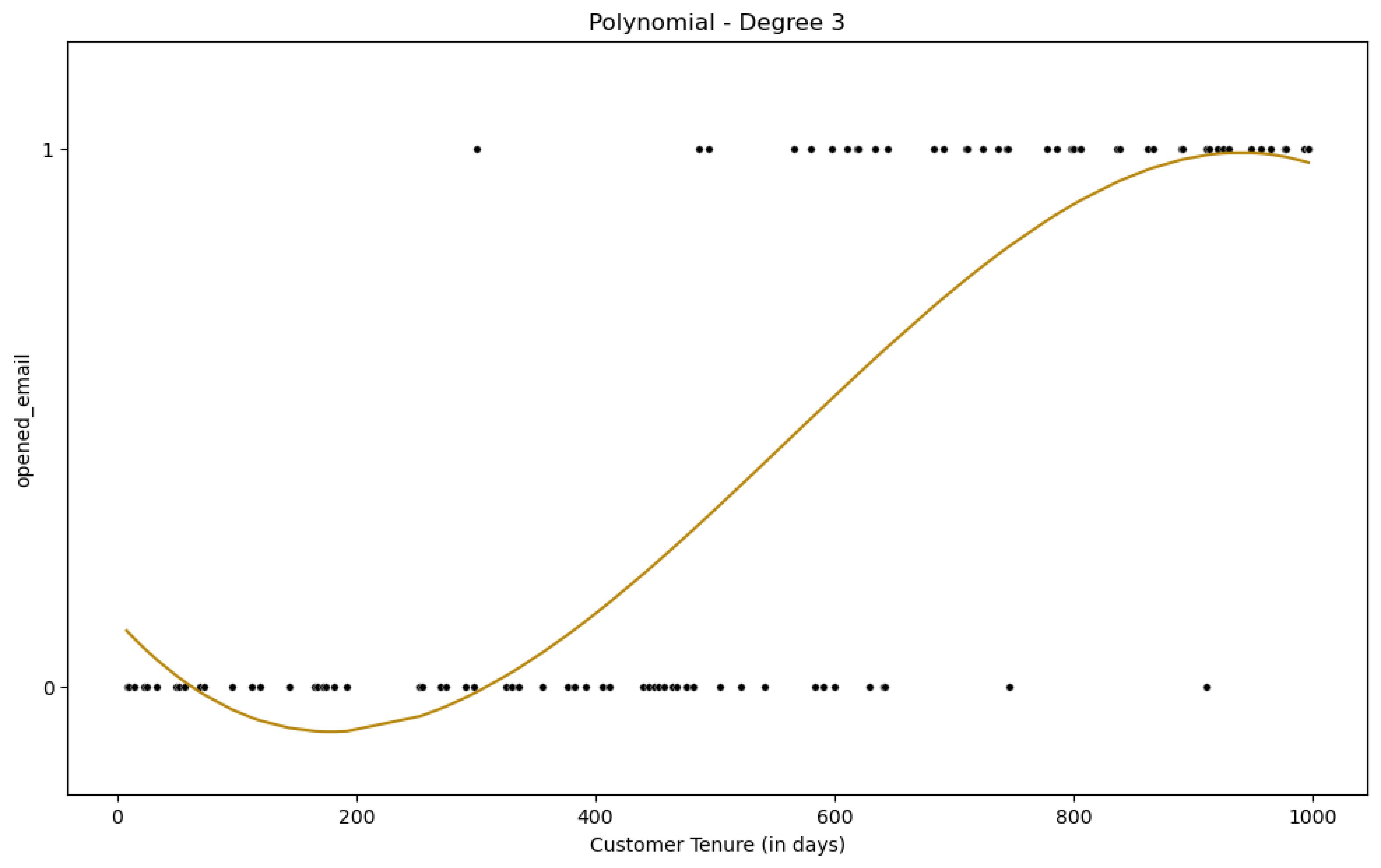
To generate a binary prediction, probabilities are converted to a binary outcome based on a cutoff value. E.g., if p > 0.6 then 1 else 0.

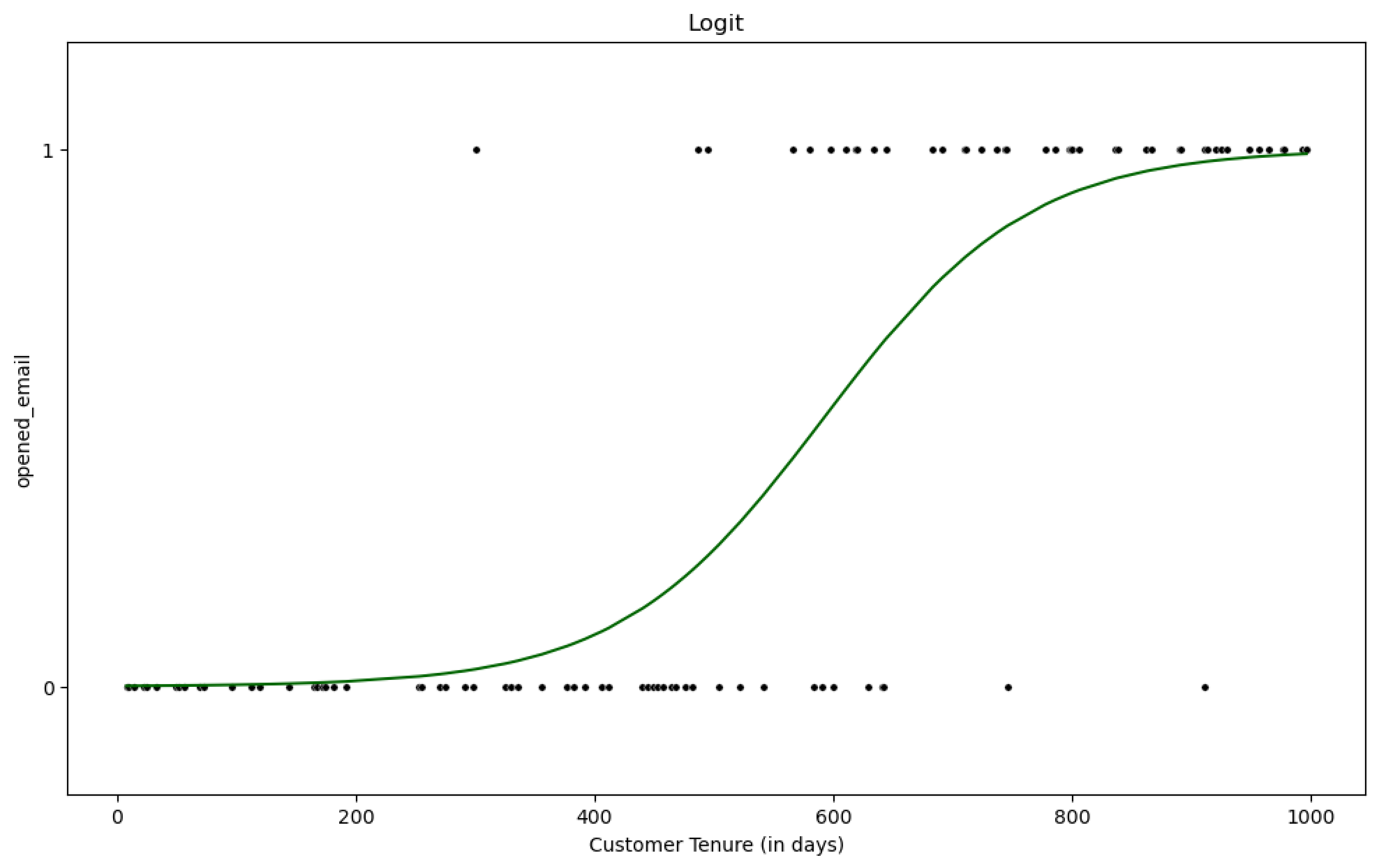
VISUALIZATION OF LOGIT FUNCTION ON BINARY DATA: LOGIT FITS BETTER


Scatterplot
1
0
opened_email
0
200
400
600
800
1000
Customer Tenure (in days)


Linear Regression
1
0
opened_email
0
200
400
600
800
1000
Customer Tenure (in days)


Polynomial - Degree 2
opened_email
1
0
0
200
400
600
800
1000
Customer Tenure (in days)


Polynomial - Degree 3
1
0
opened_email
0
200
400
600
800
1000
Customer Tenure (in days)


Logit
1
0
opened_email
0
200
400
600
800
1000
Customer Tenure (in days)

PREDICTION AND INFERENCE

Prediction

- Is there a relationship between predictor(s) and outcome
  - *Statistical test comparing fitted model to null model. Indicated by significant χ^2 difference between fitted and null model*
- How strong is the relationship?
  - *Statistical metrics derived from LogLikelihood*
    - AIC
    - Pseudo R-Squared (McFadden)
  - *Accuracy Metrics*
    - Accuracy or misclassification rate
    - Precision
    - Recall (or Sensitivity)
    - Specificity
    - Area Under the ROC Curve (AUC)

Inference

- Which predictors influence the outcome?
  - *Statistical significance of coefficient indicates relevance of predictor*
- Interpretation of coefficients?
  - *If X_i is increased by one unit, the log odds will change by b_i units, when the effect of other independent variables is held constant.*
  - *The sign of b_i will determine whether the likelihood increases (if the sign is positive) or decreases (if the sign is negative) by this amount.*

Other Issues
Class Imbalance

- When predicting categorical outcomes, a common problem is class-imbalance, where one level of the outcome is under-represented. Prediction in such cases is particularly hard because simply predicting the majority outcome will generate pretty good results. Consider the following:
  - *Predicting Customer Churn: Over a given period of time, most customers are loyal, few leave.*
  - *Predicting credit card fraud: Most credit card transactions are legitimate*
  - *Medical Imaging: Detecting an abnormal cell amongst hundreds that are healthy.*

Summary

- In this module we
  - *examined Logistic Regression*
  - *contrasted logistic regression with linear regression*
  - *learnt about the logit model and estimation*
  - *visualized logit function on binary data*
  - *examined prediction and inference*