Logistic regression considerations

Logistic Regression Assumptions

1. **Linearity**: There is a linear relationship between any continuous feature and the log(Odds) of the variable.

2. **Independence**: All values of the target are independent (each row in the dataset is independent).

Multicolinearity

- Not an assumption as in linear regression.
- It can still affect the coefficients of the regression.
- We want to remove highly collinear input features.

Problems with the data

- The Logistic regression implementations will almost always return a model with some coefficients.

- Sometimes, it will be bad, for example it will fail to converge.
  - Incomplete information
  - Complete separation

Incomplete information

Gender	Rich	Survived
Female	Yes	1
Female	No	0
Male	Yes	1

Gender	Rich	Survived
Male	No	?

Incomplete information

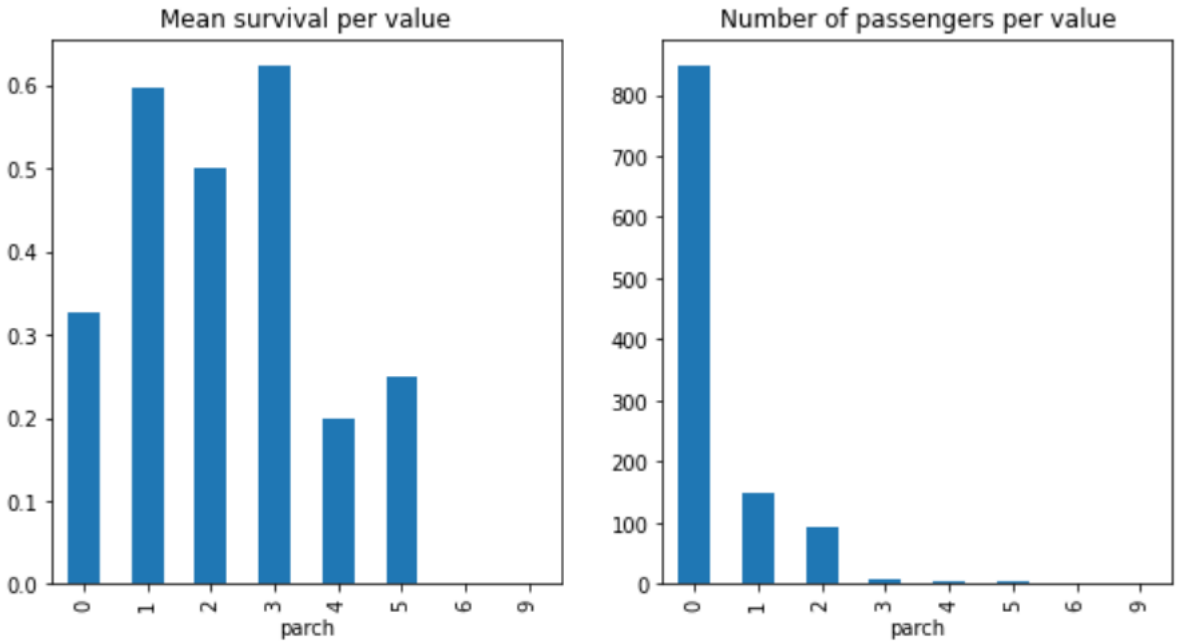
```
1 pd.crosstab(data["cabin"], data["survived"])
```

survived	0	1
cabin		
A	11	11
B	18	47
C	37	57
D	14	32
E	11	30
F	8	13
G	2	3
T	1	0

- Check that you have all options by cross-tabulation.
- Group infrequent values in categorical variables.

Incomplete information

- Check for unusual combinations.
- Check for unusual data points.
- Cap or group them.

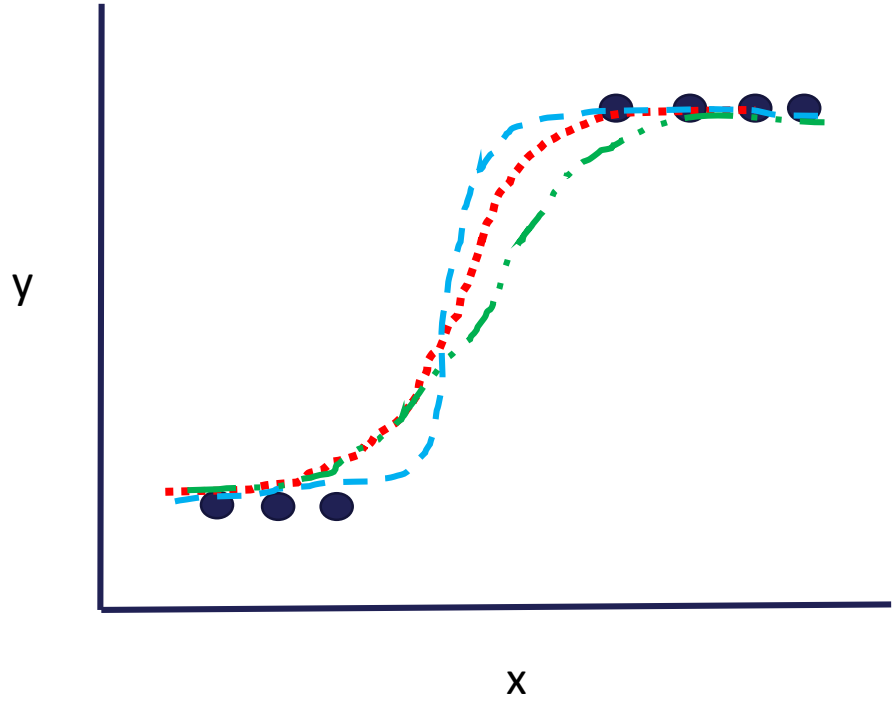


Incomplete information

- Ensure representative dataset.
- Do not trust predictions for sectors that are not well represented in your data.

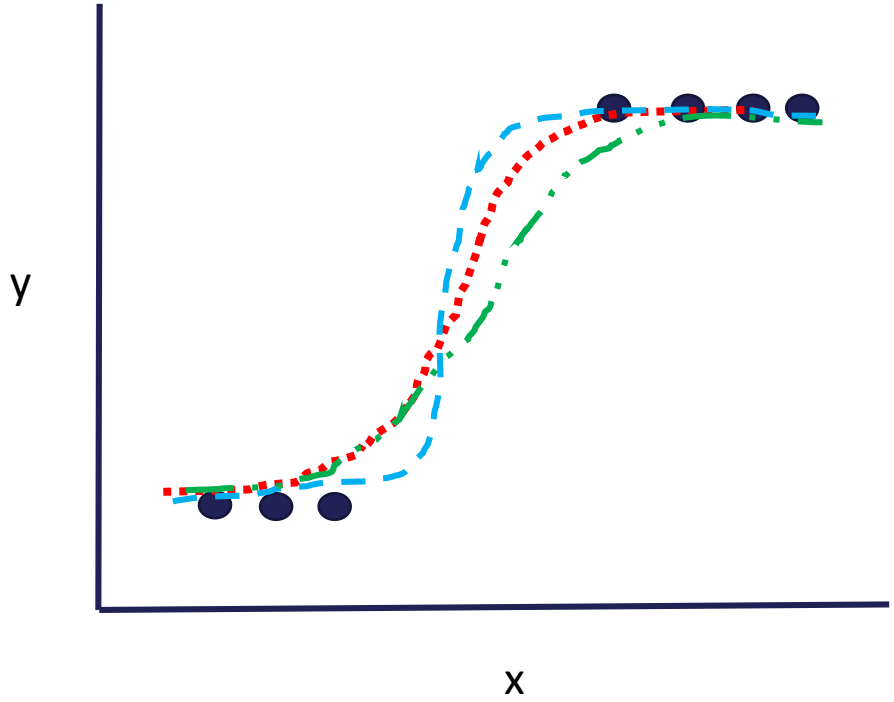
Complete separation

- The classes can be perfectly separated by 1 or a combination of variables.
- Lack of information in between classes.
- Curves will be too steep ➔ inflated errors for the coefficients.



Complete separation

- Often happens when there are too many variables and too few data points.
- Collect more data.
- Use a simpler model.



Feature Selection

- Simpler models are easier to interpret.
- Lasso can reduce the number of features by shrinking their coefficients (β_i) to zero.
- Chi2- test for categorical variables and ANOVA for continuous variables.
- Other feature selection methods (all have pros and cons)

Categorical features

- Use one hot encoding
- Group rare values.

Train In Data

THANK YOU

www.trainindata.com