Shallow Ann and Logistic Regression on Titanic data

Jason Tang

03/18/2023

C O N T E N T S

01 PART ONE Data preparation

	Sex	Age	SibSp	Parch	Fare	Title_Master	Title_Miss	Title_Mr	Title_Mrs	Title_Officer	...	Ticket_STONOQ	Ticket_SWPP	Ticket_WC	Ticket_WEP	Ticke
0	1	22.0	1	0	7.2500	0	0	1	0	0	...	0	0	0	0	
1	0	38.0	1	0	71.2833	0	0	0	1	0	...	0	0	0	0	
2	0	26.0	0	0	7.9250	0	1	0	0	0	...	0	0	0	0	
3	0	35.0	1	0	53.1000	0	0	0	1	0	...	0	0	0	0	
4	1	35.0	0	0	8.0500	0	0	1	0	0	...	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
886	1	27.0	0	0	13.0000	0	0	0	0	1	...	0	0	0	0	
887	0	19.0	0	0	30.0000	0	1	0	0	0	...	0	0	0	0	
888	0	18.0	1	2	23.4500	0	1	0	0	0	...	0	0	1	0	
889	1	26.0	0	0	30.0000	0	0	1	0	0	...	0	0	0	0	
890	1	32.0	0	0	7.7500	0	0	1	0	0	...	0	0	0	0	

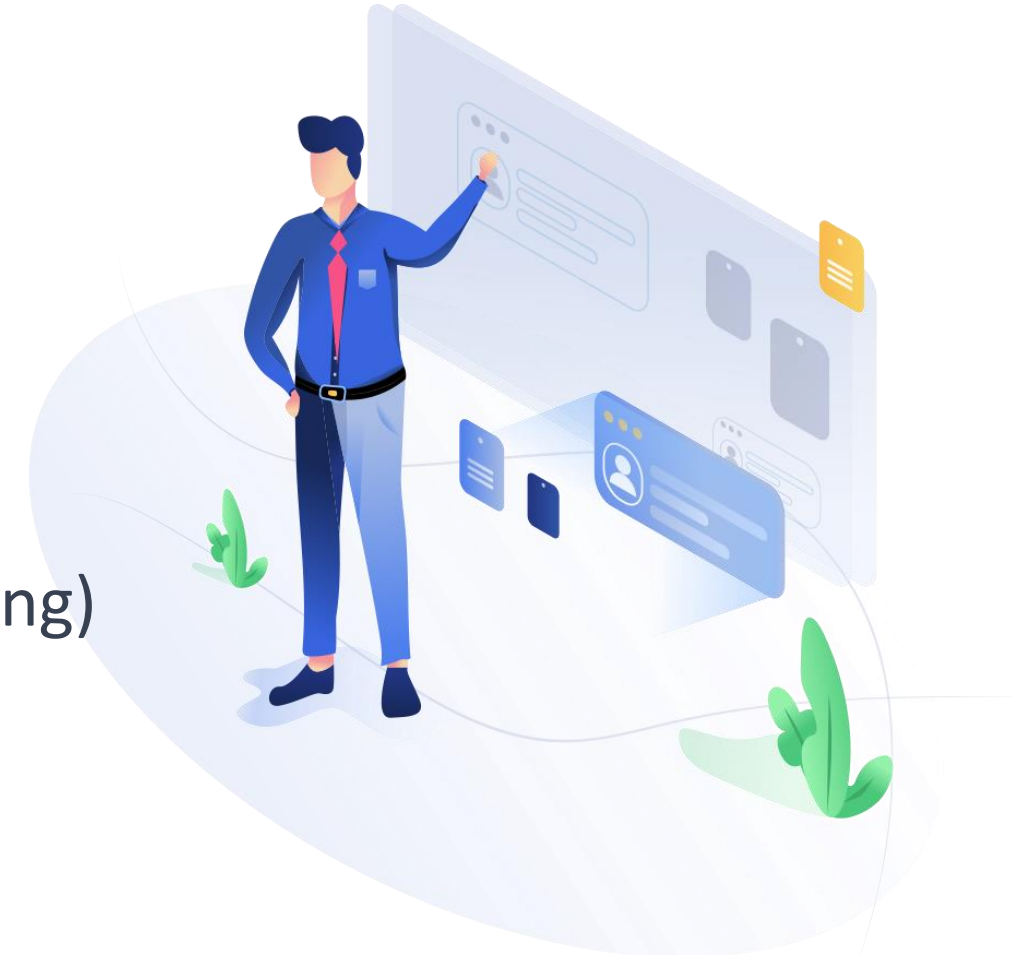
891 rows × 69 columns

```
data.isnull
```

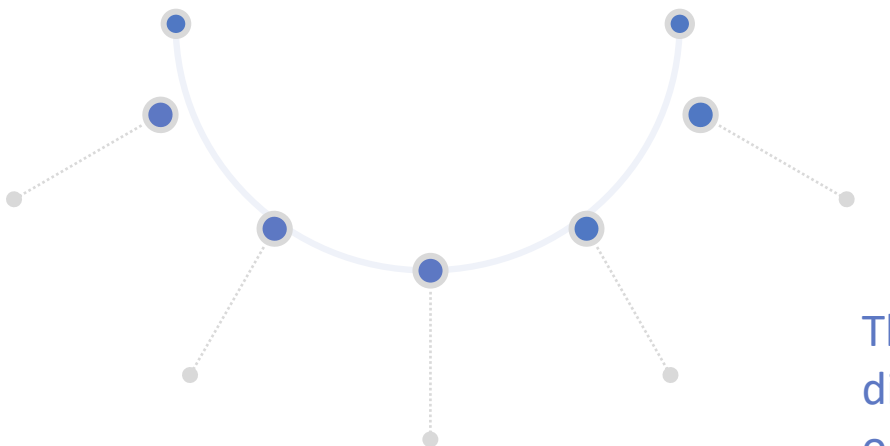
This Titanic data is after one-hot encoding and feature engineering. It contains 69 columns including the target variable "Survived". All features have been transformed into numerical format and so it is ready for model training and prediction

02 PART TWO

Feature engineering(binning)



```
# Feature Engineering
data['Age_bin'] = pd.cut(data['Age'], bins=[0, 12, 20, 40, 60, 100], labels=False)
data['Fare_bin'] = pd.qcut(data['Fare'], q=4, labels=False)
data['FamilySize_bin'] = pd.cut(data['FamilySize'], bins=[0, 1, 4, 20], labels=False)
```



The 'Age' feature is divided into 5 bins based on age ranges (0-12, 12-20, 20-40, 40-60, 60-100)

The 'Fare' feature is divided into 4 bins using the quantiles of the data

The 'FamilySize' feature is divided into 3 bins based on family size (1, 2-4, and 5 or more).

```
#  Select  Features
features  =  ['Pclass_1',  'Pclass_2',  'Pclass_3',  'Sex',  'Age_bin',  'Fare_bin',
                        'Title_Master',  'Title_Miss',  'Title_Mr',  'Title_Mrs',  'Title_Officer',  'Title_Royalty',
                        'Embarked_C',  'Embarked_Q',  'Embarked_S',  'FamilySize_bin']

return  data[features],  data['Survived']
```

Here I have selected the most important feature that is most relevant to survive in an array , and Survived in an array

I have then split the preprocessed data into training data(80%) and test data(20%)

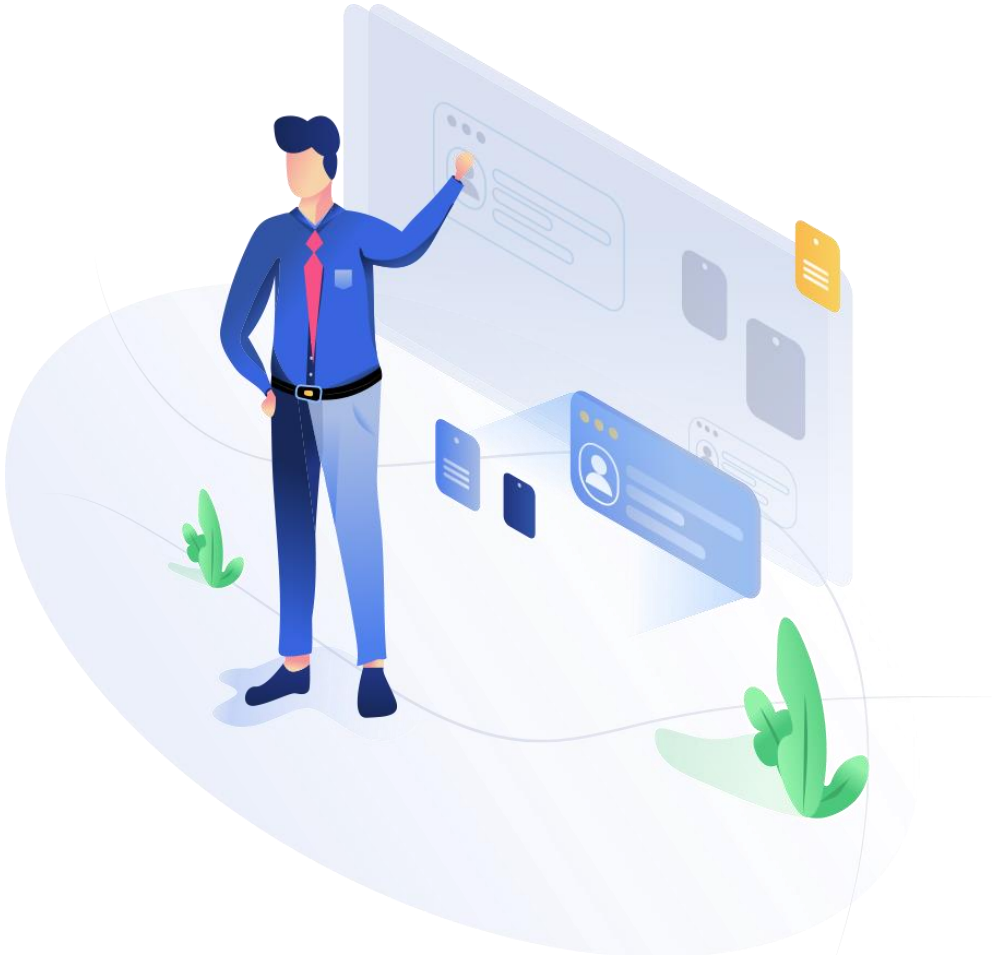
Now my data is ready for Logistic regression and Shallow ANN

03 PART THREE

Simple Logistic Regression

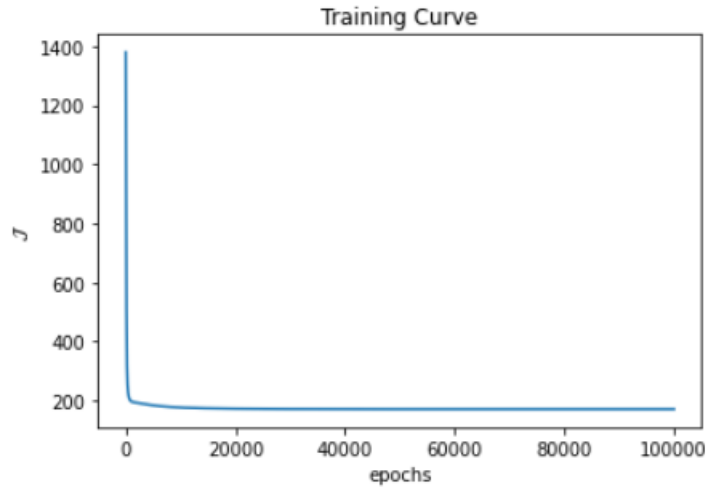
A n d

Multi-Variate Logistic Regression



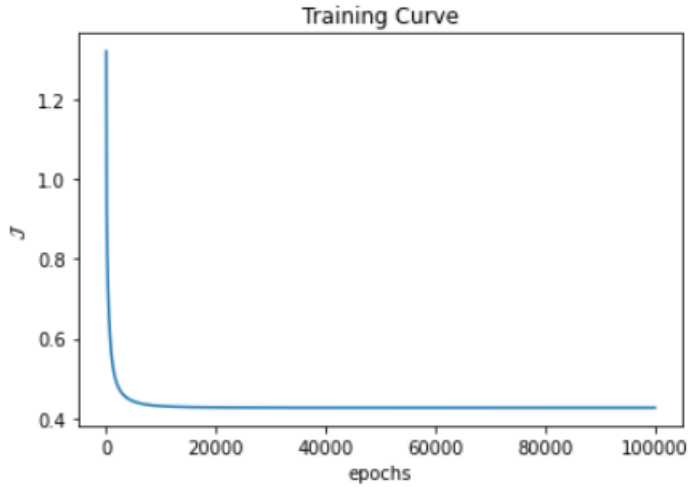
Simple Logistic regression

Test Accuracy: 0.8427



After I trained the model and evaluated the model, I got 0.8427 accuracy

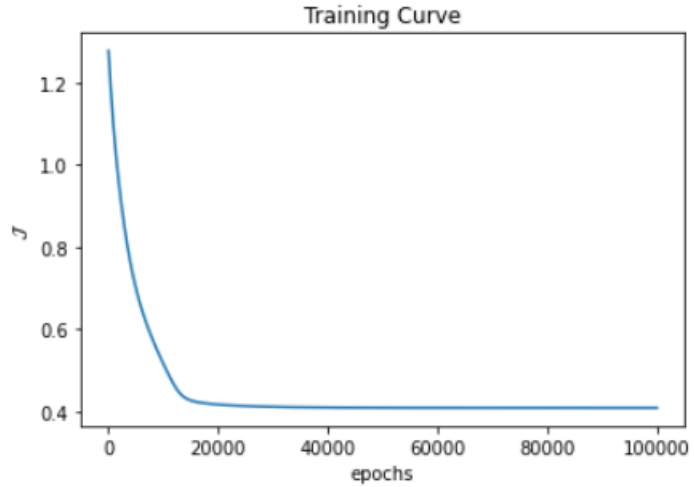
Test Accuracy: 0.8596



After some tunning and take-in more features and drop some features , the best accuracy was 0.8596

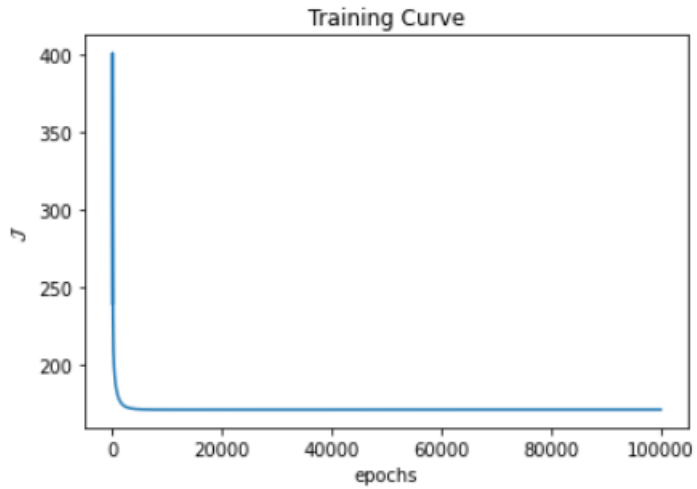
Multi-Variate Logistic Regression

Training Accuracy: 0.8418



After Tunning

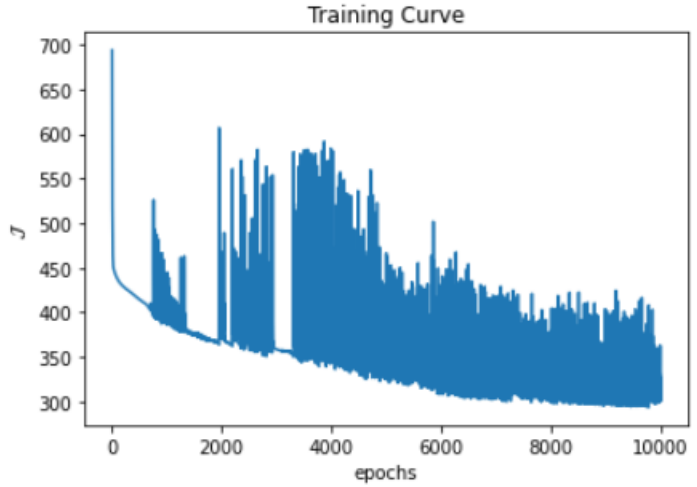
Test Accuracy: 0.8539



03 PART THREE

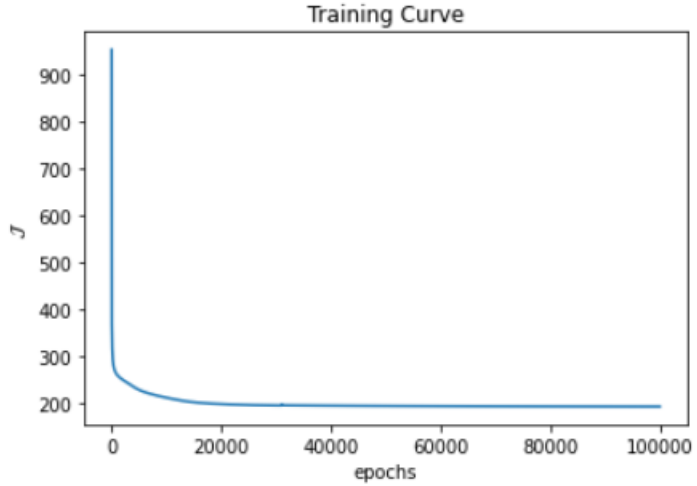
Shallow ANN

Shallow_ANN



Test Accuracy: 0.8596

I first tried without dropping any columns , and the training curve is not right here



Test Accuracy: 0.8708

After feature engineering and tunning, I have achieved a better accuracy.

Conclusion

We had a clean and ready to use dataset

We are binning age, fare and family size into specific range can be helpful in capturing potential relationships between these variables and survival.

Both Logistic and Ann worked well on titanic data set.

Shallow ANN may be a better choice for predicting survival rates in similar situations.

Demo for predicting survived or not

Thank you