Sink or Swim

Titanic Prediction Model

Alexander Garcia

December 2024

Contents

1 Problem Definition

The goal of this project is to build a Logistic Regression model that could predict whether a passenger survived the sinking of the Titanic or not. It is a binary classification problem. There are 891 samples, 12 attributes, 7 predictors, and 1 target variable. The predictor variables are Sex, Age, Pclass (passenger class), SibSp (sibling and spouse), Parch (parents and childrend), Fare, and Embarked. The target variable is Survived where 1 represents survival.

2 Background

My inspiration came from watching the movie and knowing about the Titanic since I was a kid. The Titanic is a well known event world wide and it was a good beginner data science project. The source I followed performed similar steps in terms of data cleaning that I did. However, they ran 10 different models and did more advanced hyperparameter tuning which resulted in a slightly better accuracy that I achieved[Abd].

3 Implementation Details

3.1 Dataset

As seen in the figures below some of the data is already in numerical format but others like Sex and Embarked have to be converted. Switching Sex to binary (0,1) format and then encoding Embarked where numbers 1,2, and 3 will represent the ports.

```
[7]  1 df.head(10)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C

Figure 1: Head of the data

Figure 2: Checking null value counts and balance of target variable

```
1 df.isnull().sum()
```

	0
PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	177
SibSp	0
Parch	0
Ticket	0
Fare	0
Cabin	687
Embarked	2

dtype: int64

```
1 df.Survived.value_counts()
```

Survived	count
0	549
1	342

dtype: int64

The data also has some null values in Age, Cabin, and Embarked. For Age, I checked the distribution and decided to fill in using the median as its less sensitive to outliers than mean. For Embarked it is only 2 samples so I removed them. However, for Cabin its a large portion of data so that column had to be dropped.

3.2 Correlations

Using the correlation heatmap below I was able to see which features are correlated most with my target value. Interpreting the chart I noticed Sex has the biggest correlation which does match with the "women and children" first policy the Titanic had in regards to lifeboats. The most surprising correlation (even if not strongly) is Embarked, suggesting the port people boarded did affect chances of survival.

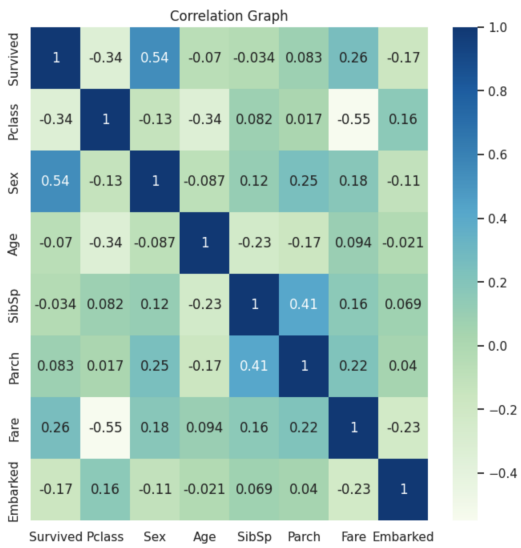

Figure 3: Correlation Heatmap

3.3 SMOTE

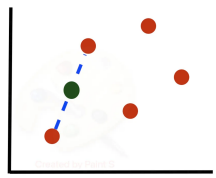

Figure 4: SMOTE source:[Mak]

Synthetic minority over-sampling technique (SMOTE) was used to address the slight imbalance for the Survived target variable. This technique generates random samples utilizing a K-nearest neighbor approach. The green dot in the chart above is the randomly generated sample that used two real samples in generation[Mak].

3.4 Logistic Regression Model

Logistic regression is good for binary outcomes like "yes" or "no", which made it a good model for my binary prediction. Unlike Linear regression this model fits an S curve to the data in order to determine the best fit and prediction.

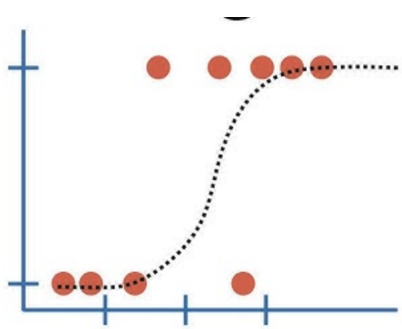

Figure 5: Source: [Sta]

4 Results

My results for training data were very close to the source model at 77.8% and 77.9% respectively. Testing accuracy was also quite close, however we differ on the F1 Score. I believe this difference is due to the source model running an automated hyperparameter tuning method known as GridSearchCV.

Figure 6: Checking null value counts and balance of target variable

	training	testing
Accuracy	0.778	0.782
F1-Score	0.775	0.778

(a) My results

```
-Train accuracy: 77.92%
-Validation accuracy: 78.98%
-Test accuracy: 81.82%
-F1_Score: 80.39%
-Precision: 83.67%
-Recall: 77.36%
```

(b) Source results [Abd]

5 Conclusion

In summary, I successfully ran a logistic model that had results very close to my source model even though I did not perform the same hyperparameter tuning. This model did not score the best validation accuracy out of all the models my source ran, but it was a good model and dataset to learn from. It had a good amount of beginner friendly tasks such as mutating Sex to binary and encoding the Embarked variable as well as working with synthetic sample generation.

References

[Abd] Walid Abd-elhameed. *Apply 10 models to Titanic Dataset*. URL: https://www.kaggle.com/code/walidabdelhameed/apply-10-models-to-titanic-dataset?scriptVersionId=196557170&cellId=43.

[Mak] Cory Maklin. *Synthetic Minority Over-sampling TEchnique(SMOTE)*. URL: https://medium.com/@corymaklin/synthetic-minority-over-sampling-technique-smote-7d419696b88c.

[Sta] Josh Starmer. *StatQuest: Logistic Regression*. URL: https://www.youtube.com/watch?v=yIYKR4sgzI8.