Predictive analysis on Titanic data

Viktoria Meszaros

2021

Topic

I chose a really interesting topic for my analysis. It was the tragedy of the luxury steamship called Titanic. As most of you already know, Titanic was one of the first ocean liners and the largest ship in the world at that time. It was traveling from Southampton to New York when an unexpected accident happened. It bumped into an iceberg and sank in the early hours of April 15, 1912. This was one of the deadliest maritime catastrophes of all time leading to the death of more than 1500 people (out of the 2224 on board).

Aim of the research

In this project my aim will be to build a predictive model that helps to find out what sort of people were more likely to survive the sinking of the Titanic than others. If there are some factors that increased the probability for someone to survive.

Data collection

The data I used for my research I found on the website Kaggle.com . It contains 819 observations (and 418 in a test file). This is because this data was uploaded as part of a machine learning competition where it is needed to have a train and than a test set. I could not join these two samples, as for the test data the binary values of if the passenger survived or not were missing, so I would not be able to build a model on that. On the Titanic there were 2435 people when it sunk, out of which 1320 were passengers and the remaining 892 crew members. In our data set we have data about 819 passengers which is more than 60% of the total population. We know that our sample was created in a totally random manner so we can assume that it is representative. There were some missing values in my data for some observation representing 8% of our data. Most of the missing values are in Cabin numbers and Age.

Variable	Number_of_NAs
Cabin	687
Embarked	2
Age	177

As Age is an important variable for my analysis I did some imputation to replace NA values. As the distribution of age was skewed to the right instead of mean value I used the median to decrease the bias. I calculated the different median age values for males and females in different social groups as I found out from the data that women were younger on Titanic and people in the 1st class were older on average compared to the ones in lower social classes. The following table shows the different values I imputed for the different groups.

Sex	Pclass	average age
female	1	35
female	2	28
female	3	22
male	1	40
male	2	30
male	3	25

Although this imputation helps to get rid of missing values it also causes bias in my estimation as these values are just my assumptions instead of actual values. According to data quality there is one more issue to mention. There is a variable called Pclass which shows the ticket class a given passenger had. From this we will assume that these passengers belong to that socio-economic layer. This variable is interesting as out initial assumption is that people who belong to the 1st class will have a higher probability of survival as they were more important is the eyes of crew during the evacuation process. This assumption may not be true if for example a really wealthy person from social class 1 bought a 2nd or 3rd class ticket as he/she does not care about luxury during his/her travel. We should also keep in mind the this tragedy happened in 1912 when administration was not even close to perfect and due to this our data may contain some measurement errors due to personal mistakes or lost documents.

Data cleaning

During the data cleaning process I dealt with missing values. I excluded Cabin variable as 77% of the values were missing. For age as I already outlined I made imputation as I considered it an important variable for my prediction. Last but not least for Embarked there were 2 missing values. I decided to do some extra research as I knew the name of the passengers. I found out that both of them embarked in Southhampton, so I filled in this values which solved the problem. After dealing with missing values I also looked at extreme values. I only had two continuous variables Age and Fare, so I only had to look for extremes here. For Age the distribution looked close to normal with no unexpected extreme values. For Fare the distribution was rather skewed to the right, due to some extreme values. I had a closer look on these, and realized these were not errors but the most expensive tickets so I decided not to exclude any values. I also had to do some transformations on categorical variables to make them appropriate for the modeling. I had some data about how many parent, siblings, spouses or children a passenger had on board with them. From this I decided to create a binary variable Travel_alone which was was if the person was alone on board and 0 if he/she had some family members with them. I though this variable will be helpful as it was said that during the rescue families, especially women or men with children had higher priority to get on the boats. For social class and embark location I created dummy variables with n-1 categories always leaving out one category as a base. For more information about my data cleaning process please check my data_cleaning.R available in my Github repository.

For my model I will use:

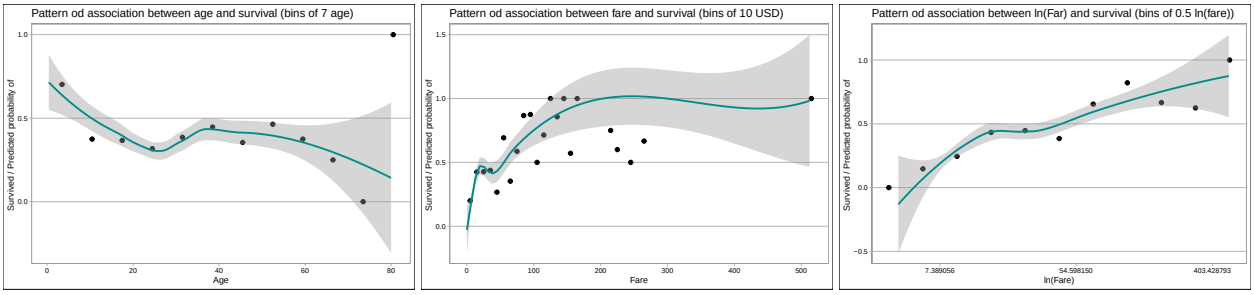
- Survived (0 if no, 1 if yes) as the dependent(y) variable
- Sex, Age, Social Class, Embarked, Ticket price and if someone traveled alone or with family are going to be the right hand side(x) variables for my model

Patterns of association

After looking at all the patterns of association between my outcome variable and all the explanatory variables I could come up with some interesting assumptions. (For all the graphs of the patterns of association please look at the Appendix section)

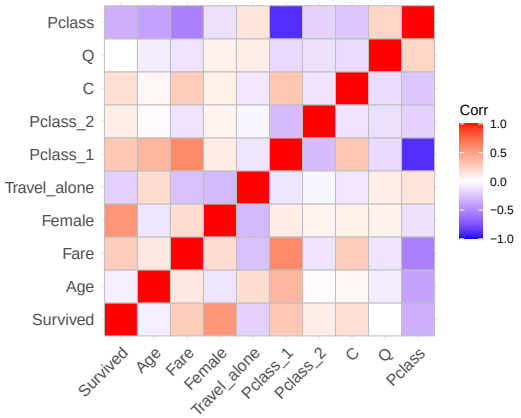
- Women had more chance to survived then men.
- People in higher social class had higher probability of survival. Someone in socio-economic class 3 had the worst chance.
- Those who embarked in Southhampton died with higher probability and those who got on board in Cherboug had the highest chance of surviving.
- Someone who traveled with his/her family also were in a better situation as for them survival was more likely compared to someone traveling alone.
- The probability to survive decreased with age for people below 25 and above 32, but increased between these two
- As the price people paid for the travel increased, so did the probability of survival

From the graphs below you can see the patterns of association between the two continuous variables and the dependent variable. Based on this I decided to use a piecewise linear spline for Age and the log transformation for the Fare variable later on.



Correlation between variables

Before building my model I checked the correlation between my variables. From this graph you can see that there is only one variable pair whose correlation is high and that is Pclass and Pclass_1. This is what we expect as they measure the same thing. We have moderate correlation between Pclass and Fare as Pclass measures the class of the ticket its price should increase if it belongs to a higher class.Also there is moderate correlation between Female and Survived meaning that the sex of the passenger strongly correlated with the probability of survival. For our model the only correlation that matters is between the fare and Pclass. Due to this I decided not to use fare in my model only social class. This also made my final model simpler, with keeping its power.



Model formula

As I did prediction I decided to split my data to two sets. First to a training set and then a test set to do robustness check in the end. I put randomly 80% of my observations to train and left the remaining 20% in a test set.

I ran my predictive models on the train set. I tried to predict if someone survived the sinking of Titanic or not when I know their sex, if they traveled alone or with their families, their social class (1st, 2nd or 3rd), where they embarked (Queenstown = Q, Cherbourg = C or Southhampton = left-out category) and their age.

$$P(Survived)^P = Female + Travelalone + Pclass_1 + Pclass_2 + Q + C + lspline(Age, c(27, 32))$$

For this model formula I ran a liner probability model, a logit and a probit model to find the one that gives the best results. This table shows the coefficients for the **LPM** model (*Model 1*) and the marginal effects for the **logit** (*Model 2*) and **probit** (*Model 3*) models.

	Model 1	Model 2	Model 3
Constant	0.282**		
	(0.062)		
Female	0.496**	0.491**	0.494**
	(0.032)	(0.035)	(0.035)
Pclass_1	0.341**	0.344**	0.334**
	(0.041)	(0.045)	(0.045)
Pclass_2	0.181**	0.176**	0.167**
	(0.039)	(0.035)	(0.035)
lspline(Age, c(27, 32))1	-0.011**	-0.010**	-0.010**
	(0.003)	(0.003)	(0.003)
lspline(Age, c(27, 32))2	0.016	0.016	0.017
	(0.009)	(0.010)	(0.009)
lspline(Age, c(27, 32))3	-0.007**	-0.007**	-0.007**
	(0.002)	(0.002)	(0.002)
Travel_alone	0.052	0.046	0.046
	(0.032)	(0.029)	(0.029)
Q	0.121*	0.115*	0.110*
	(0.052)	(0.050)	(0.050)
C	0.114**	0.112**	0.116**
	(0.038)	(0.040)	(0.041)
Num.Obs.	712	712	712

* $p < 0.05$, ** $p < 0.01$

Model interpretation

We can see that the values are really similar for all of the variables meaning that the functional form of my explanatory variables seems to be good.

- With 99% confidence we can assume that the probability for a women to survive is 50% higher than for a men
- With 99% confidence a person belonging to the 1st social class has 34% higher probability to survive compared to someone from the 3rd social class, while someone in the 2nd class has 18% higher chance
- We can be 99% confident that under the age of 27 and over the age of 32 the probability of survival is expected to be 1% lower for someone one year older

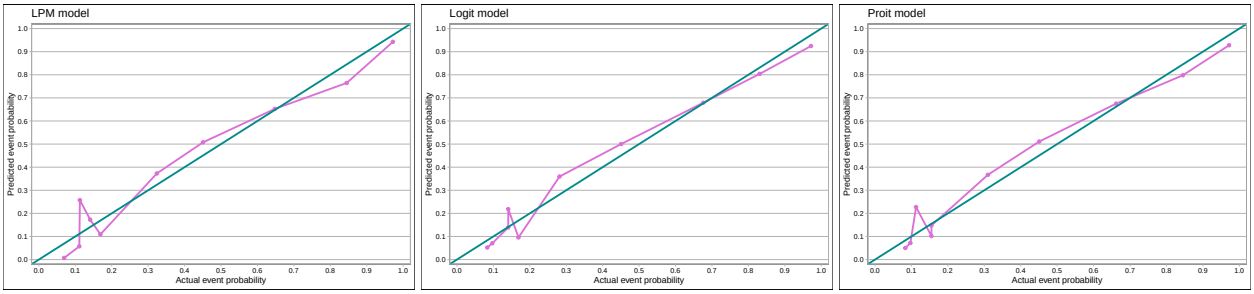
- Also there is 99% confidence that those who embarked in Cherbourg has 11,5% more probability to survive compared to those who embarked in Southhampton. And with 95% confidence we can also state that people embarked in Queenstown had 11-12% more probability to survive than who got on board in Southhampton
- We cannot be confident that traveling alone had any effect on the probability of survival

Final model

When deciding between LPM, logit and probit models I looked at the goodness of fit with BIC, checked model fit with brier and also made some model diagnostics like bias and calibration curves. We can see that the BIC scores are similar but it is the smallest for the logit model. Considering the bias values LPM and logit seems to be unbiased (0.000 bias score values) on the other hand for probit we get 0.003 which is low as well but compared to the others really high. The Brier score is also similar, but it is the lowest for the logit model. Anf last but not least the prediction accuracy is 80% fro the LPM while 81% for logit and probit models.

Stats	LPM	Probit	Logit
logLik	-315.610	-314.860	-313.710
AIC	653.230	649.730	647.410
BIC	703.480	695.410	693.090
Bias	0.000	0.003	0.000
Brier-Score	0.142	0.139	0.138
Prediction Accuracy (%)	79.900	80.900	80.900

From the graphs below we can also conclude that all of the models are quite well calibrated, as the predicted probabilities tend to move along the 45° line on the y ~ yˆ curve.



Due to these reasons and values I picked the logit model as my final choice as that is the best according to most of the metrics.

Analyse predicted probabilities

Highest probability of survival From the model we could find out that the lukiest ones who had the highest probabbility of survival were on average 24 years old women belonging the the 1st socio-economic class, who embarked in Cherbourg and travelled alone travelled alone.

statistics	Female	Travel_alone	Pclass_1	Pclass_2	C	Q	Age
mean	1	0.5714286	1	0	0.8571429	0	24.71429
median	1	1.0000000	1	0	1.0000000	0	32.00000
sd	0	0.5345225	0	0	0.3779645	0	13.18730

Lowest probability of survival On the other hand the worst chances belonged to men with an age around 54 coming from the lowest socio-economic class, embarking in Southhampton who travelleda alone.

statistics	Female	Travel_alone	Pclass_1	Pclass_2	C	Q	Age
mean	0	0.75000	0	0	0	0.1250000	54.00000
median	0	1.00000	0	0	0	0.0000000	53.25000
sd	0	0.46291	0	0	0	0.3535534	14.86367

Robustness check

From the comparison of the coefficient and the marginal effect of out models we can say that the functional form I chose for the explanatory variables seems to be good. In my analysis.R file you can also check that I built my final model formula through several steps including my variables step by step. The coefficents/marginal effects are close to constant throughout my models showing that they actually has an effect of the specific size on the outcome variable. For robustness check I also ran the final model on my test data set. From the confusion table we can see that the model has similar accuracy as for the training data set with 77% of prediction accuracy. I am satisfied with this result, I think this shows that my model is quite robust.

	0	1
0	0.5251397	0.1452514
1	0.0893855	0.2402235

External validity

I think the external validity of my prediction is really low. It could be used to predict the survival of passengers on Titanic, but not on other boats. Here most of the findings that they saved women and children and people who belonged to higher social classes may not be relevant in the world we have today. I would also say that there are better explanatory variables to predict survival in a catastrophe like this for example distance to rescue ships or health of a passenger. The aim of this analysis was to find out if I can build a model that predicts death in this very situation rather than to create a tool for other, later use.

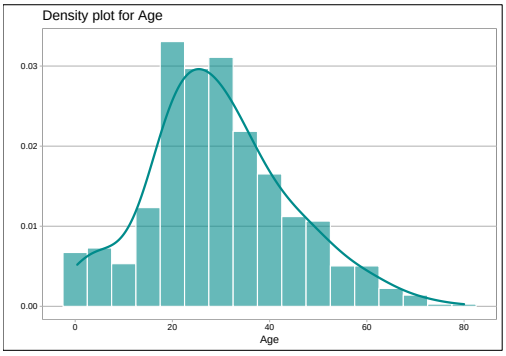
Conclusions

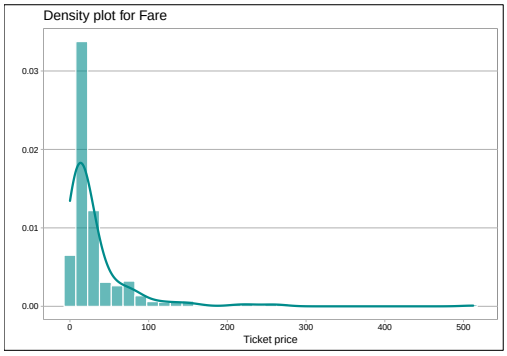
The most important findings of my analysis was that on Titanic during its sinking young females with an average age of 24 from the highest social class had the highest probability of survival. This is most probably because during the evacuation they prioritized women over men and higher social classes had also priority compared to lower ones. What surprised me is that those who embarked in Cherbourg had also significantly higher probability of survival. This may due to that the cabins were assigned by embarking and those who embarked at the first stop Southhampton got cabins on lower levels that those who got on the boat in Cherbourg. Another reason can be that more 1st class people embarked there. Overall I think these are pretty interesting findings andare close to the reality. I also managed to build a model on a random traning set (with 81% predicting accuracy) with which I could predict 77% of my test sample corretly. This shows to me that my model is highly valid for other sets of passengers on Titanic, but probably not really valid for other cases.

APPENDIX

I. Data descriptives

Continuous variables

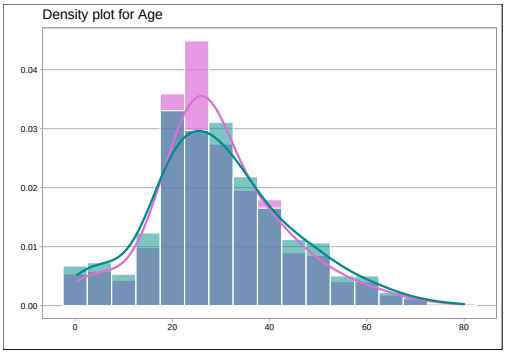




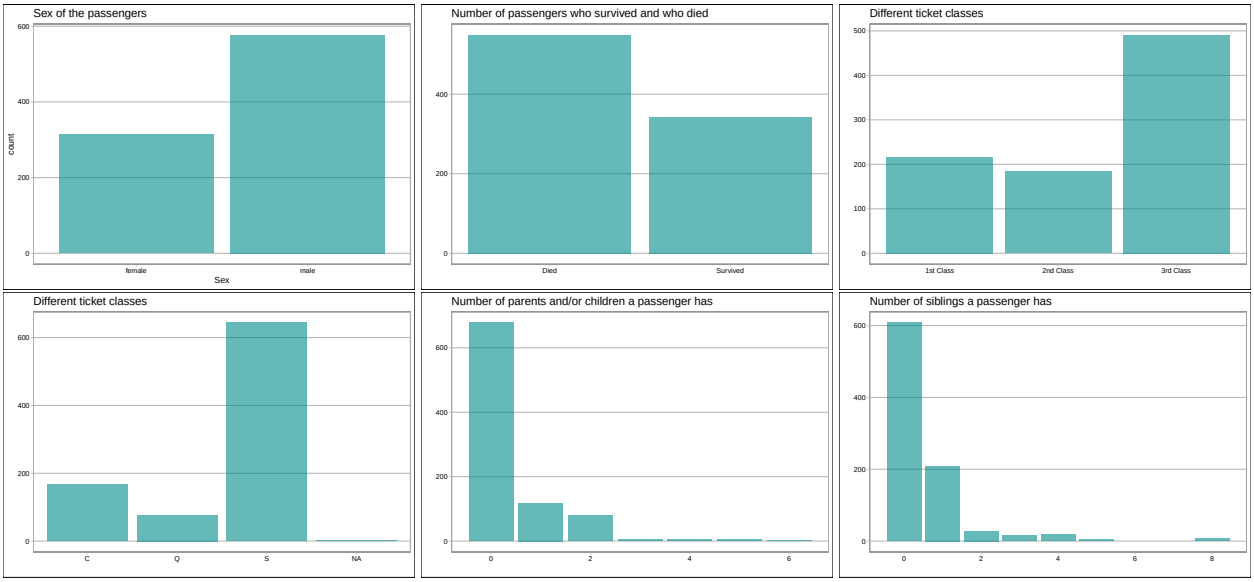
Summary statistics for continuous variables

statistics	Age	Fare
min	0.42000	0.00000
1st_qu.	20.12500	7.91040
median	28.00000	14.45420
mean	29.69912	32.20421
3rd_qu	38.00000	31.00000
max	80.00000	512.32920
sd	14.52650	49.69343
# missing	177.00000	0.00000
# used obs	714.00000	891.00000

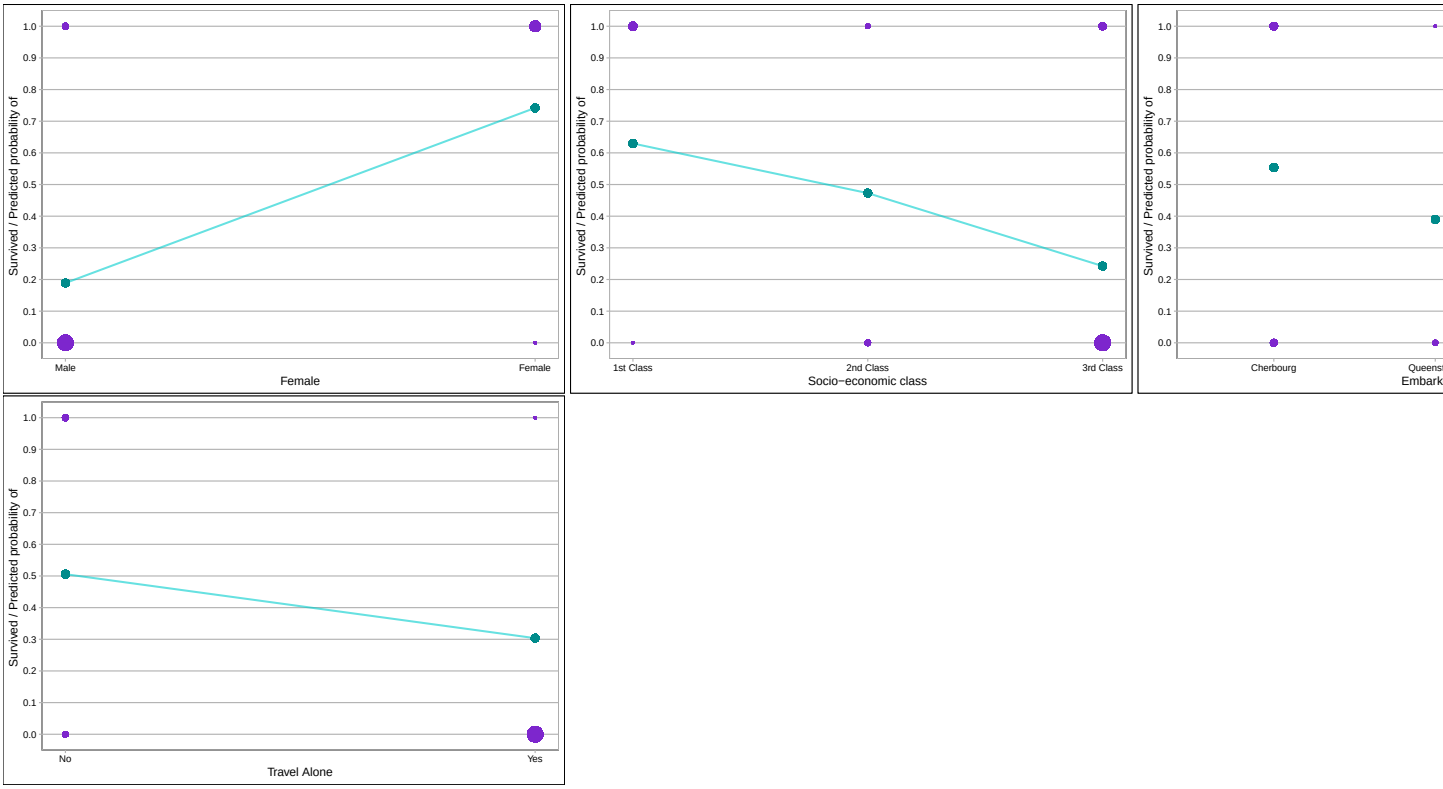
Distribution of Age after imputation



Categorical variables

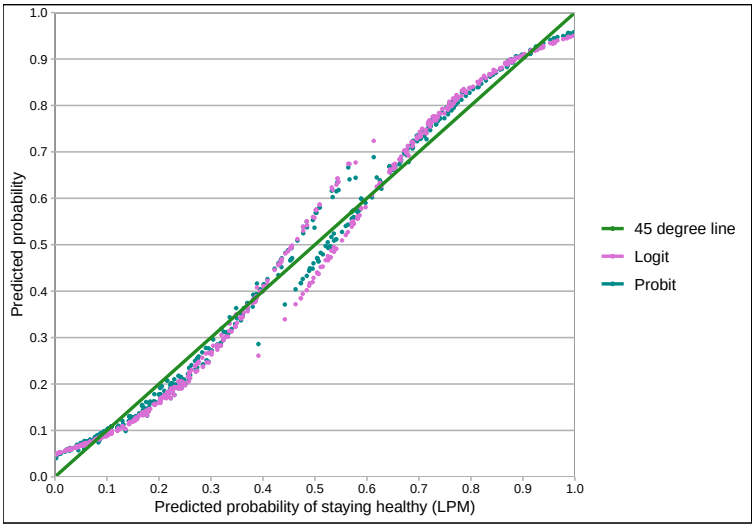


II. Patterns of association

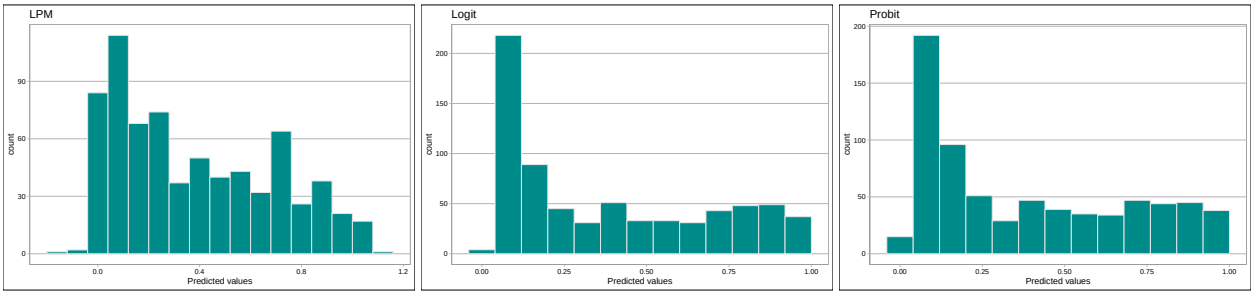


III. Modelling

Comparing predicted probabilities of logit and probit to LPM



Predicted values by models



Confusion tables

```
##
##               0         1
##   0 0.52808989 0.11235955
##   1 0.08848315 0.27106742
##
##              0         1
##   0 0.5421348 0.1165730
##   1 0.0744382 0.2668539
##
##               0         1
##   0 0.53792135 0.11235955
##   1 0.07865169 0.27106742
```

Summary statistics for all models where Survived == 1

statistics	pred_lpm	pred_logit	pred_probit
mean	0.6294316	0.6364231	0.6372544
median	0.7026648	0.7339342	0.7285307
min	-0.0003341	0.0489852	0.0463489
max	1.0600412	0.9667463	0.9762815
sd	0.2788640	0.2821090	0.2794601

Summary statistics for all models where Survived == 0

statistics	pred_lpm	pred_logit	pred_probit
mean	0.2304446	0.2260968	0.2311348
median	0.1594352	0.1282379	0.1307715
min	-0.1803318	0.0147326	0.0075423
max	1.0967857	0.9714296	0.9773189
sd	0.2094321	0.2040077	0.2050560