This significance test will check to see if the data provides convincing evidence that the F1 score achieved by the model is statistically greater than the average F1 score obtained by guessing.

If conditions are met, the test will use a 1 sample t test for proportions (F1 score is a proportion of precision and recall). The conditions are as follows: a random sample, normality, and large population size. The null and alternative hypothesis must also be stated. The following variables are used in the checking of the conditions and in later calculations:

1. : the population proportion (in this case, an F1 score of 0.5)
2. : the sample size (in this case, there are 267 prices used in the testing data)
3. : the sample proportion (in this case, the model’s F1 score of 0.5926)
4. ; the z statistic of the test
5. ; the z-score value in a normal distribution
6. ; the null hypothesis
7. ; the alternative hypothesis

Random

* This can be disregarded in this instance, since it is necessary that the model predict on the most recent days of data

Normality

* Requirements: and .
* , this requirement is satisfied
* , this requirement is satisfied

Large Population

* Requirements:
* Bitcoin existed for over 2920 days > 10 \* 267, this requirement is satisfied

Hypothesis



Since all conditions are met, calculations can now begin.

Now that we have calculated our p value (0.0012), the results can be stated.

The p value is 0.0028. Since the p value, 0.0012, is less than , the null hypothesis is rejected.

There is sufficient evidence to show that the F1 score of the model is statistically greater than the F1 score obtained by guessing. If the proportions really were the same, then the probability of getting a sample F1 score as extreme as this is 0.0012. This is statistically significant at the significance level, and any other reasonable significance level (i.e. ). The null hypothesis, is rejected. This means that the idea that the model obtained an F1 score of .5926 through guessing is rejected.

Since the null hypothesis has been rejected, there is the probability of making a Type I error. This would be if the idea that the model obtained such a high F1 score by change is rejected, when in reality it is true. However, since it was tested at the significance level, the probability of making this type of mistake is only (or 1%), since and .