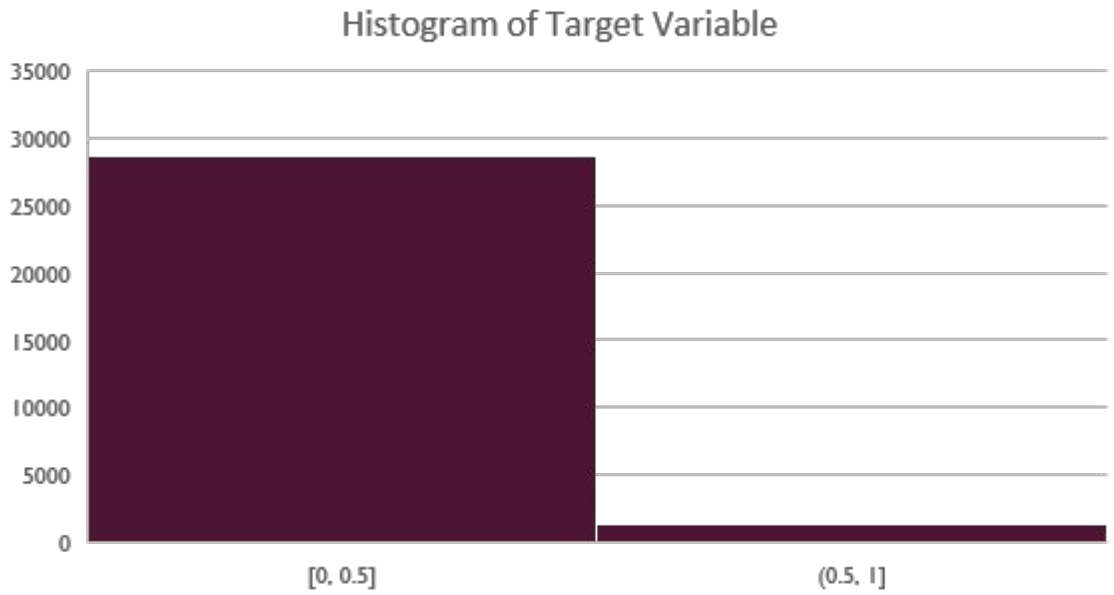
BINARY RESPONSE VARIABLE MODELS

The Problem

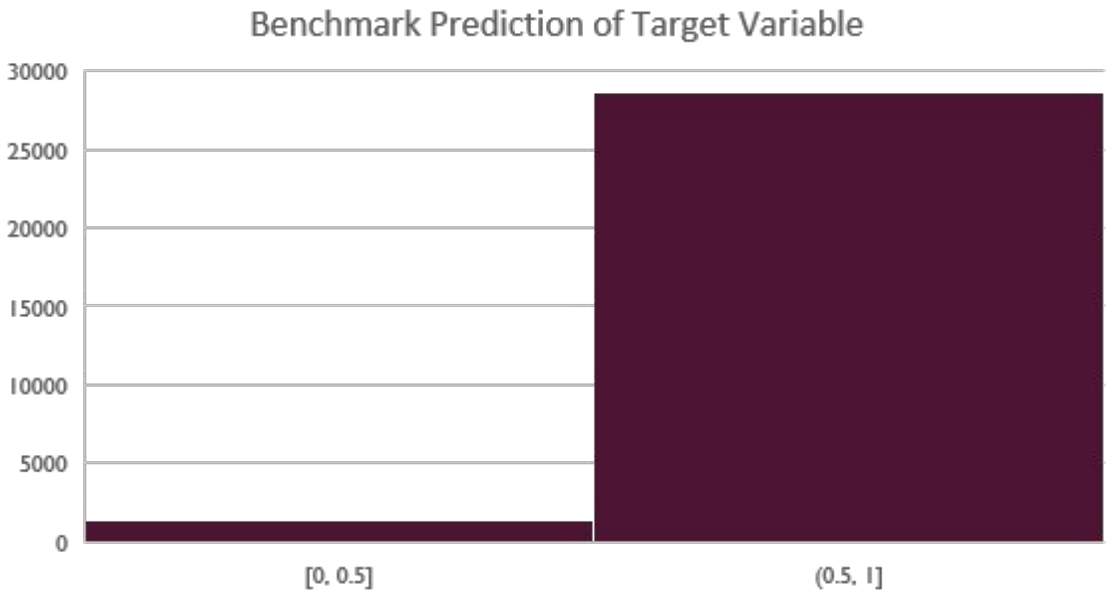
- A dataset with 30,000 customers and 50 undefined variables has a binary response variable.
- A response variable is the variable we wish to predict, and the value is either 0 or 1.
- The response variable is only 1 in 4.36% of instances, which means the dataset is imbalanced.



BINARY RESPONSE VARIABLE MODELS

The Problem

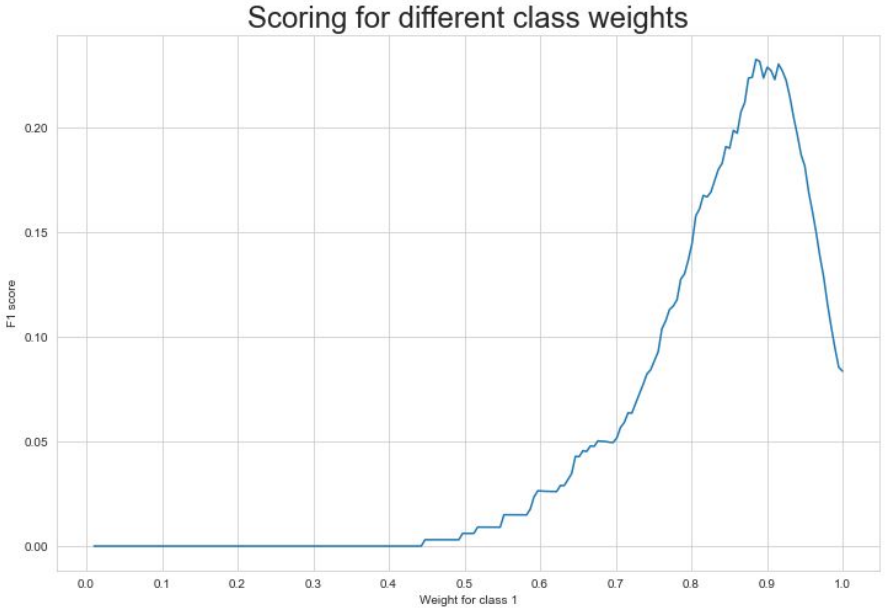
- A dataset with 30,000 customers and 50 undefined variables has a binary response variable.
- A response variable is the variable we wish to predict, and the value is either 0 or 1.
- The response variable is only 1 in 4.36% of instances, which means the dataset is imbalanced.
- We want to use logistic regression and another model to see if we can improve upon a precalculated benchmark, which has many false positives.



BINARY RESPONSE VARIABLE MODELS

The Approach

- Logistic regression is commonly used for binary response variables.
- The output of a logistic regression model is a probability between 0 and 1.
- We were able to tune the weights of the model in order to reduce the number of false positives.
- A random forest model was also created, but that model was not suitable.



A grid-search of weights were used to calculate F1 scores for the model.

BINARY RESPONSE VARIABLE MODELS

Results and Further Thoughts

- Using all variables in a logistic regression and L2 error, we were able to greatly reduce the false positive rate.
- Client and project needs dictate whether false positives or false negatives are preferrable, and which class weights are best.
- The dataset did not include variable names. Additional data cleaning, variable selection, and model selection could use information related to the variables.

Predicted Values	Actual Values: 1	Actual Values: 0
1	True Positive 360 (18.53% of positive predictions)	False Positive 1583 (81.47% of positive predictions)
0	False Negative 948 (3.38% of negative predictions)	True Negative 27109 (96.62% of negative predictions)