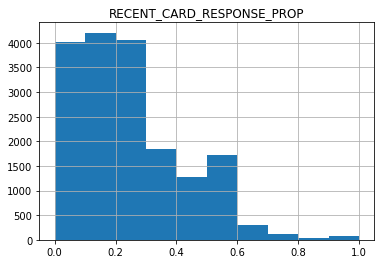
1) Create a Histogram for a continuous variable



a) Provide the mean, median, standard deviation, and confidence intervals.

Mean = 0.23072644290735914, median =0.2, std =0.1843329918196034

95% Conf\_interval of mean = [0.22799461390180295, 0.2334522621612428]

95% conf\_interval\_med = [0.2, 0.2]

95% conf\_interval\_std = [0.18216858890375917, 0.18648396842674786]

b) Explain what these descriptive statistics tell us about the variable distribution

Mean and Median give us central tendency of the proportion of card responses,

the standard deviation gives us the dispersion of the proportion of card responses.

Confidence intervals of the measures give us a 95% confidence that the population parameters are within these bounds

c) Does the variable follow a normal distribution? Explain your answer.

Variable doesn't appear to follow normal distribution given asymmetry and positive skew

2)Create a Correlation matrix for all continuous variables and Chi-square test of association for all categorical variables -please see jupyter notebook

a) Provide examples of the different ways variables may or may not be correlated, and explain.

Most of the variables weren’t strongly correlated, however months since first gift and months since origin were strongly positively correlated; the longer someone’s been in the system, the longer ago their first donation was made.

Recent card response proportion was moderately negatively correlated with month’s since last response, which intuitively makes sense (the longer since last response, the lower their proportion of responses would be).

b) Provide examples of the different ways variables may or may not be associated, and explain.

I seemed to have a lot of dependent categorical variables. I used 95th percentile for test statistic to determine association. Donor Gender and Wealth Rating appeared independent, while Wealth Rating and Urbanacity appeared highly dependent

3) Build a Linear Regression Model using a target and predictor variables

a) Provide detailed explanation of the results as it relates to the model fit, statistical measure of the

variables, and model assumptions. - Attempted to predict Target D, model was not extremely well fit, given error measures:

Mean Absolute Error: 5.651818582171258

Mean Squared Error: 62.69302381140112

Root Mean Squared Error: 7.917892637021616 (this was within 1 std, although not normal dist)

Assumed non-normality and non constant variance so data was scaled during preprocessing.

b) Provide suggestions to either improve the model’s prediction or provide a new approach to the

Prediction. - oversample from positive class for training. Keep holdout as is.

4) Compare the Linear Regression Model versus other Machine Learning Methods

a) Train the same data and target with 3 other machine learning methods of your choice

-Please see notebook

b) Compare all models based on their results. Which is the best model? I chose a RF regressor(w 10-fold cross val and grid search), Decision Tree regressor, and Support Vector Regressor, in addition to the linear model. The results were not great, the linear model actually performed the best, but I think that’s because I need to do further tuning of the non linear models. I used MSE or negative MSE when cross validated to compare models. All scores are in notebook.

c) Explain which model statistics support the best model and why? The linear model had the lowest mean squared error. The support vector regressor was close, and actually had an improved MAE over the linear model but a worse MSE/RootMSE (which penalize larger errors more than MAE)

5) Build the Best Classification Model using Machine Learning Methods

a) Train the same data using the binary target with at least 5 different machine learning methods of

your choice -pls see notebook, target B was used for target

b) How would you go about comparing the accuracy of each? I used f1 score along w confusion matrix, given imbalance.

c) Which method provides the best accuracy based on the comparison? What were the accuracy

measures you used to support the best the model? The highest f1 score was for decision tree at 0.28, with KNN and RF ~0.26. SVM and SGD did poorly, but these model parameters were not necessarily optimized.

d) Explain why you think that method performed the best? I think decision tree did well because it’s non-linear, and I don’t think this data was very separable (also why KNN did well)