Summary of Credit Card Project

Derek Baker

02 Nov 2022

derek.clayton.baker@gmail.com

Overview

- This slide show summarizes a two-part project I completed for improving my skills at using Python for data science.

- The project uses a dataset downloaded from Kaggle, containing customer information and behavior for 30,000 Taiwanese credit card users.

- The project is divided into two parts. **Part I** focuses on building machine learning models; **Part II** on running statistical tests and simulations.

Table of Contents

About the Data

- The dataset was downloaded from Kaggle.
- It contains anonymized information for 30,000 Taiwanese credit card users from April to September 2005.
- According to the information provided at Kaggle, the data was uploaded from the UC Irvine Machine Learning repository.

About the Data

- There are multiple ways in which **the data is questionable.**
- Most importantly:

- Over 22% of customers in the sample default.
- The average customer has a debt of approximately 1.4x their credit limit.
- The roughly 70% of customers have either exceeded their credit limit or missed at least one payment.

About the Data

- For these reasons, I think the data does not represent a random sample of customers of the credit card.

- However, it may represent a **random sample of customers who have been flagged as at risk of defaulting**. This would explain some of the oddities.

About the Data

- My working assumption is that it is a random sample of flagged customers.

- But in real life, I would want more information on this data, and how the customers were selected for the dataset, before drawing any conclusions about how the data can be used in supporting business decisions.

About the Data

- A final problem with the data:

- Customers sometimes jump from being zero months late in their credit card payments, to being several months late.

- I.e., a customer might be 0 months behind in payments in June, but in July they are recorded as being 2 months behind.

About the Data

- I cannot tell if this is a mistake.

- Dropping the data, or “correcting” it (e.g., changing the records so the customer is only 1 month late rather than 2) does not improve machine learning performance.
- The models seem, if anything, to perform slightly worse after making these corrections.

Part I – Machine Learning Models

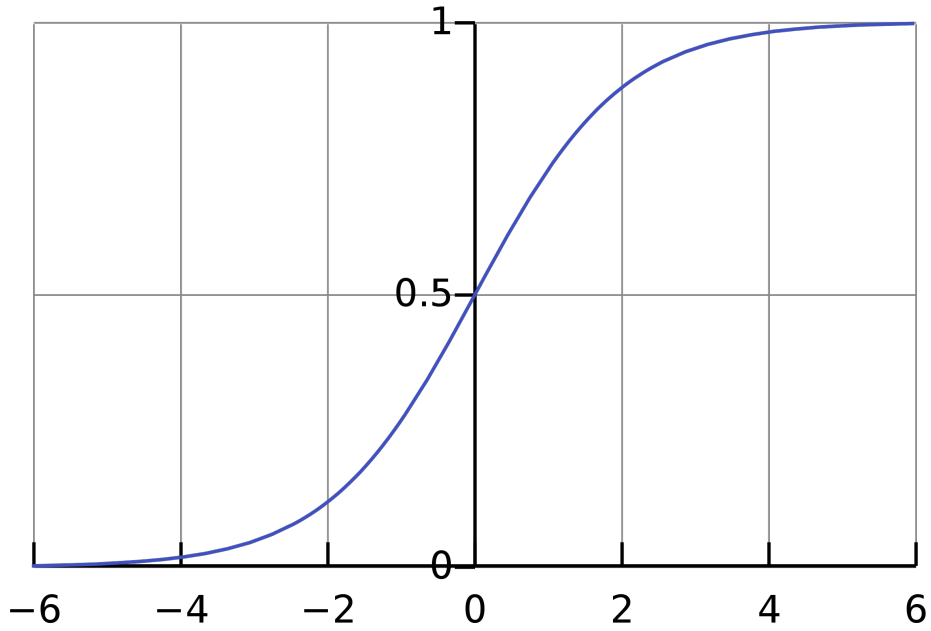
- Part I of the project focuses on training machine learning models to predict which customers are likely to default.

- This is a classification problem with imbalanced data (22.2% default, 77.8% do not default).

- Features are both quantitative and categorical.

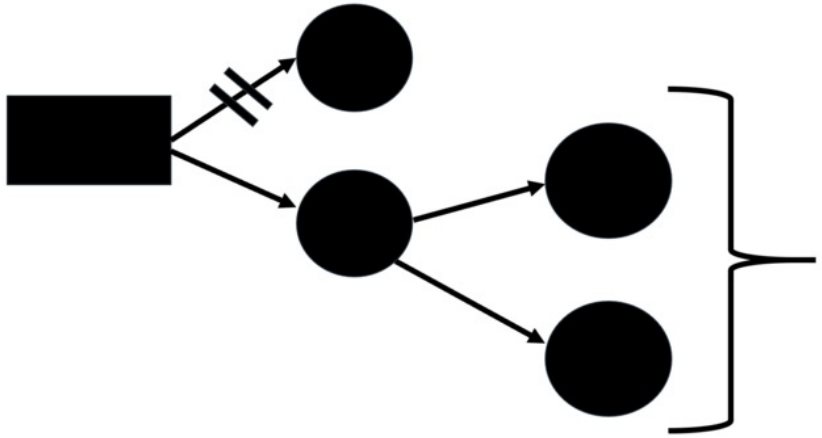
Part I – Machine Learning Models

- Two kinds of models were used.
- A linear model – Logistic Regression



Part I – Machine Learning Models

- Two kinds of models were used.
- And a non-linear model – Random Forest



Part I – Machine Learning Models

- Models were measured primarily on two metrics.
- Predictive Performance, which was measured using the area under the ROC curve (ROC AUC).
- This seemed to provide metric given that the target data was imbalanced.

Part I – Machine Learning Models

- Additionally, ROC AUC measures the tradeoff between False Positives and True Positives.

- A company may decide that it will accept more false positives to ensure more true positives (if the cost of a customer defaulting is high).

- A company may also decide to miss cases of default to avoid false positives (if the cost of misidentifying a customer as likely to default is high).

Part I – Machine Learning Models

- Besides performance, models were also ranked in terms of interpretability.

- Can we easily state, in more or less ordinary English, what the models are doing?

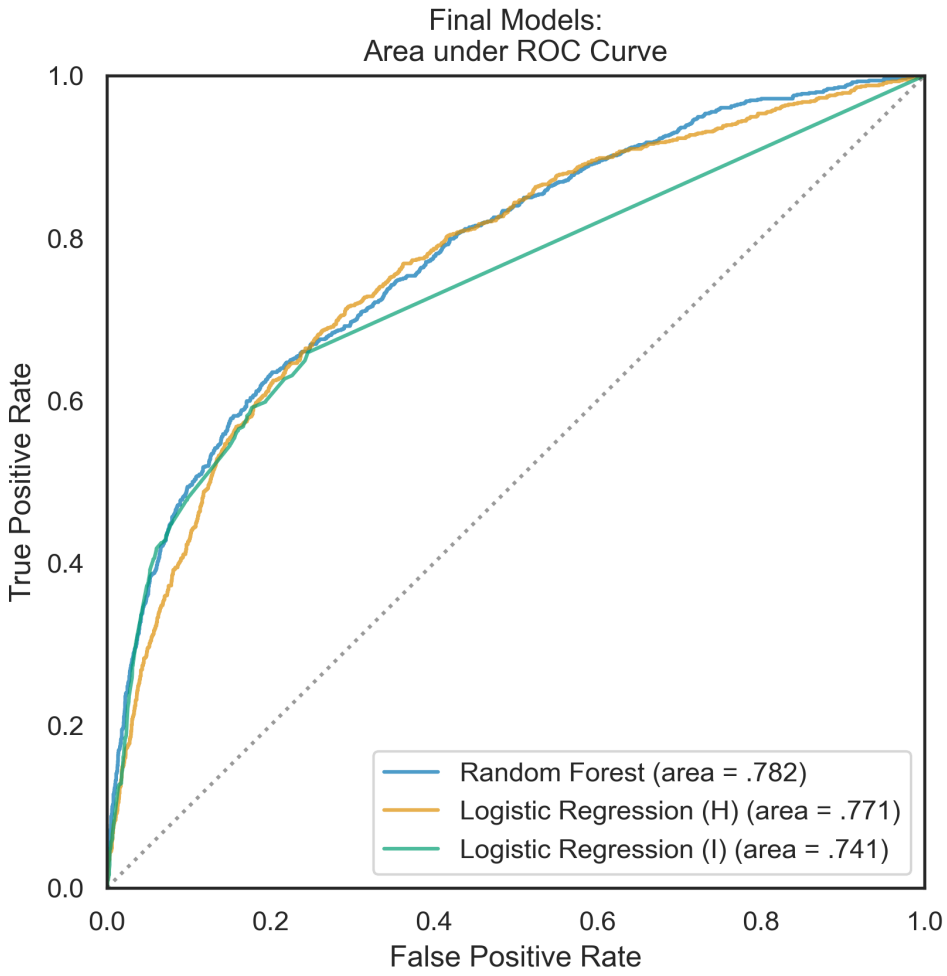
Part I – Summary of Results

- The data was divided into training, testing, and validation sets.
- Testing data was used to doublecheck cross-validation scores while tuning the models.
- This double-checking was unnecessary in some cases, but in others feature engineering involved scaling or PCA decomposition.
- A separate testing batch was held out to protect against data leakage.

Part I – Summary of Results

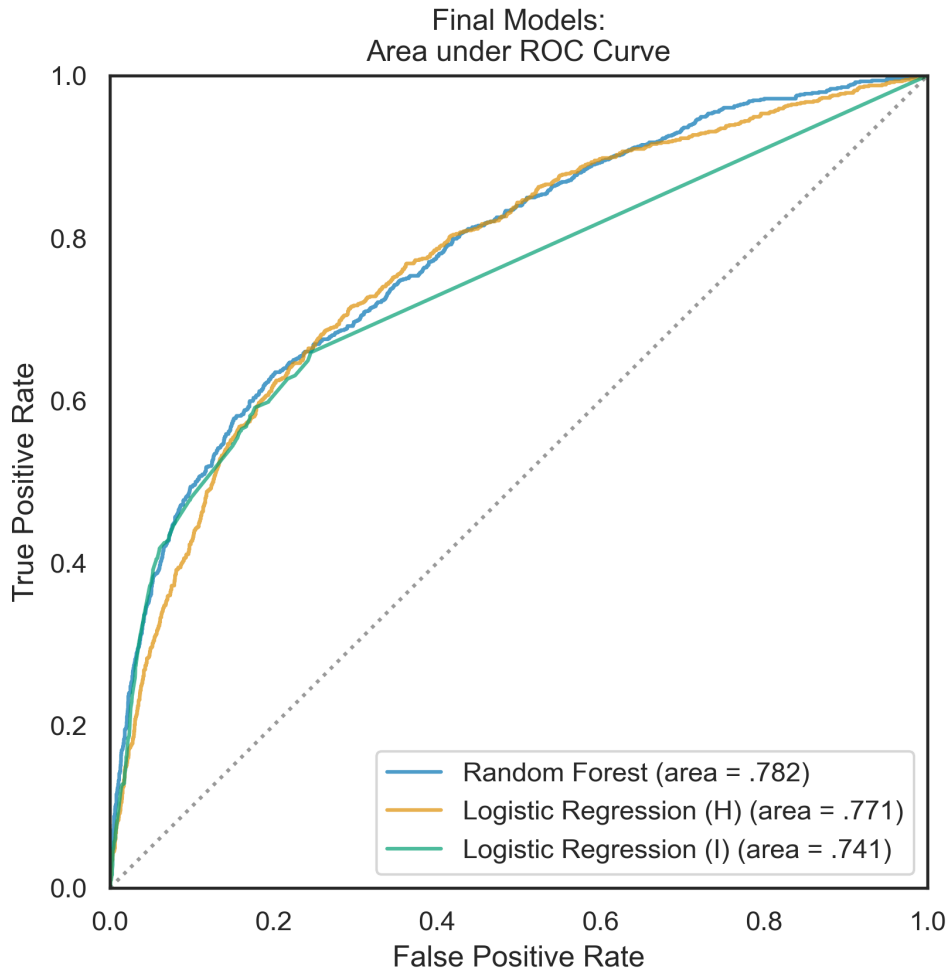
- There were three final models tested on the validation set.

- A Random Forest Model had the best predictive performance (ROC AUC = .782).



Part I – Summary of Results

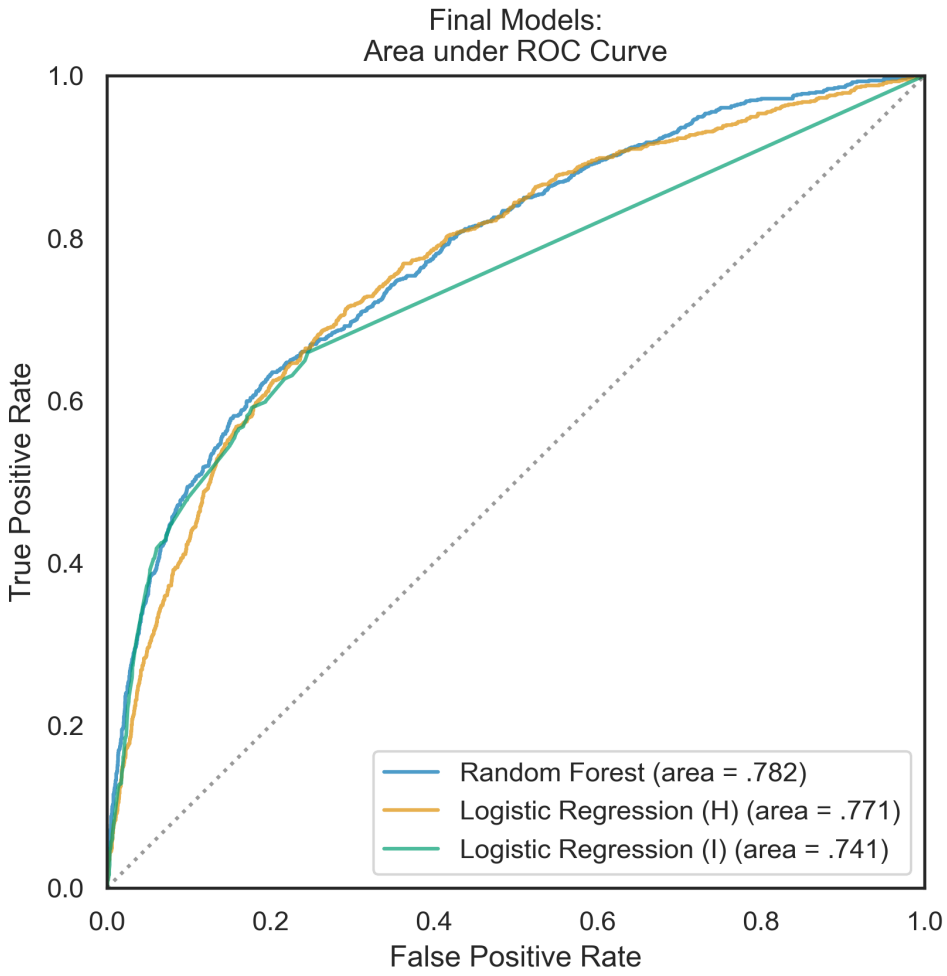
- A logistic regression model had an ROC AUC of .771.
- This model used substantial amounts of feature engineering (normalizing data, scaling, and PCA decomposition).



Part I – Summary of Results

- Logistic regression model H (for “heavy) had an ROC AUC of .771.

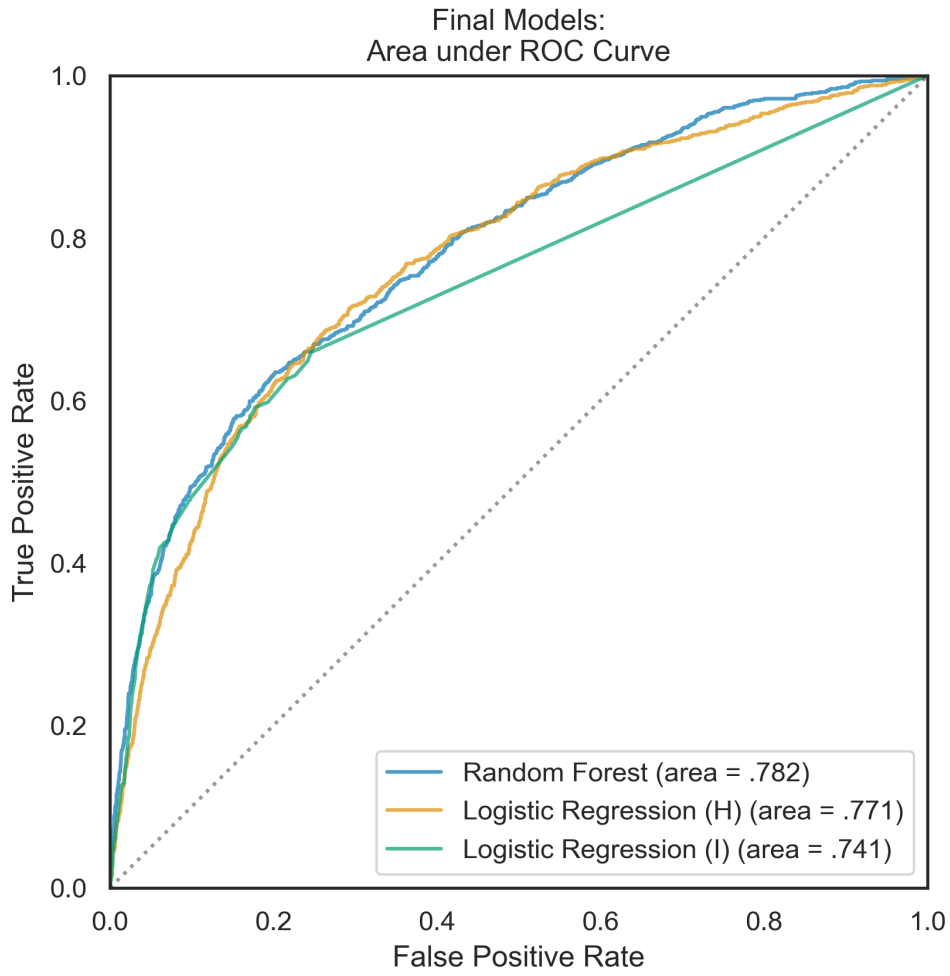
- This model was trained on data that was normalized, scaled, and decomposed with PCA.
- This improved model performance, at price of interpretability.



Part I – Summary of Results

- Logistic Regression model I (for “interpretable”) had the weakest performance.

- But it was also the most easily interpreted model.



Part I – Summary of Results

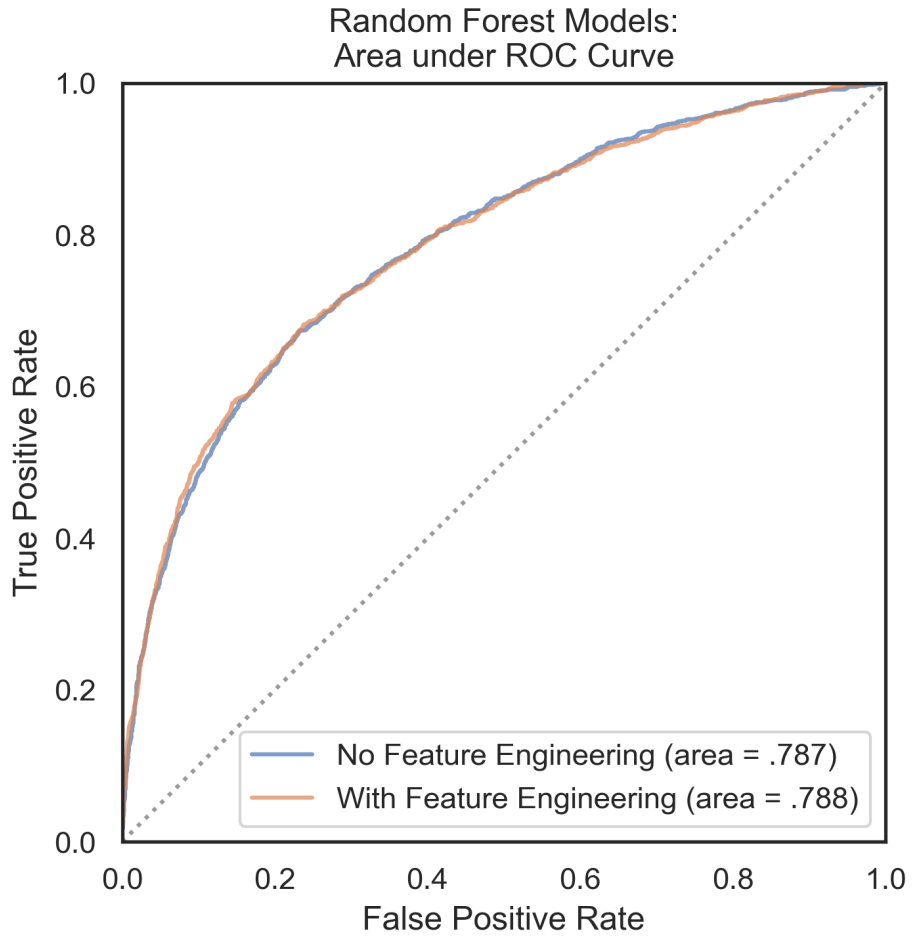
- The best performing logistic regression models used a balanced class weight.
- This improved ROC AUC; but hurt raw accuracy.
- The best random forest model did not use weighting to balance the data, so it also has a higher accuracy score.

Part I – Summary of Test and Validation Scores

	ROC_AUC_test	ROC_AUC_valid	Accuracy_test	Accuracy_valid
Random Forest	0.788	0.782	0.819	0.821
Logistic Regression - (Heavy Processing)	0.772	0.772	0.753	0.751
Logistic Regression - (Easy to Read)	0.725	0.741	0.821	0.819

Part I – Random Forest

- The random forest model did not require much work in terms of feature engineering.
- Random Forest scores before and after feature engineering were nearly identical.
- This shows scores on the test (not validation) set.



Part I – Random Forest

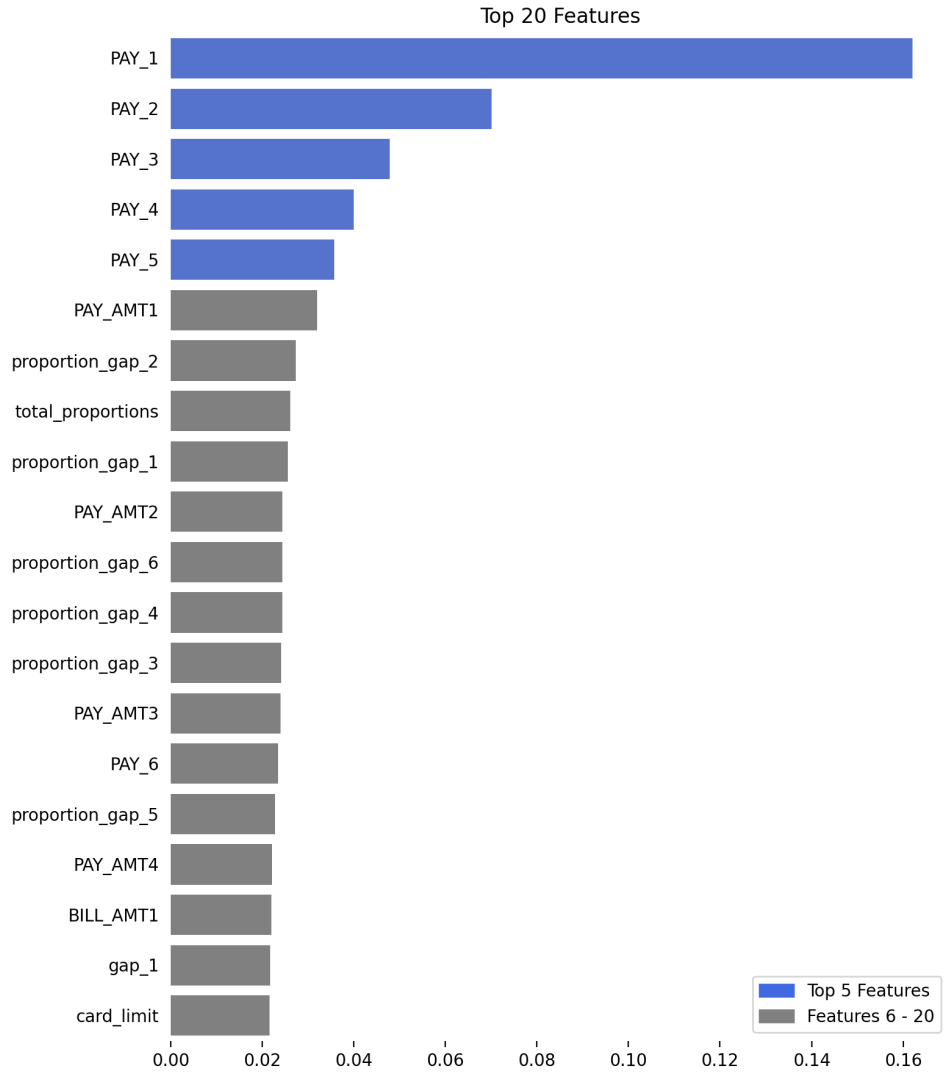
- The random forest model has the best ROC tradeoff, the highest accuracy score, and requires very little in the way of feature engineering.

- It is, however, notably slower than the linear model.

- With a larger dataset logistic regression would be preferable; or else we would need to train the forest on a random subsample of the dataset.

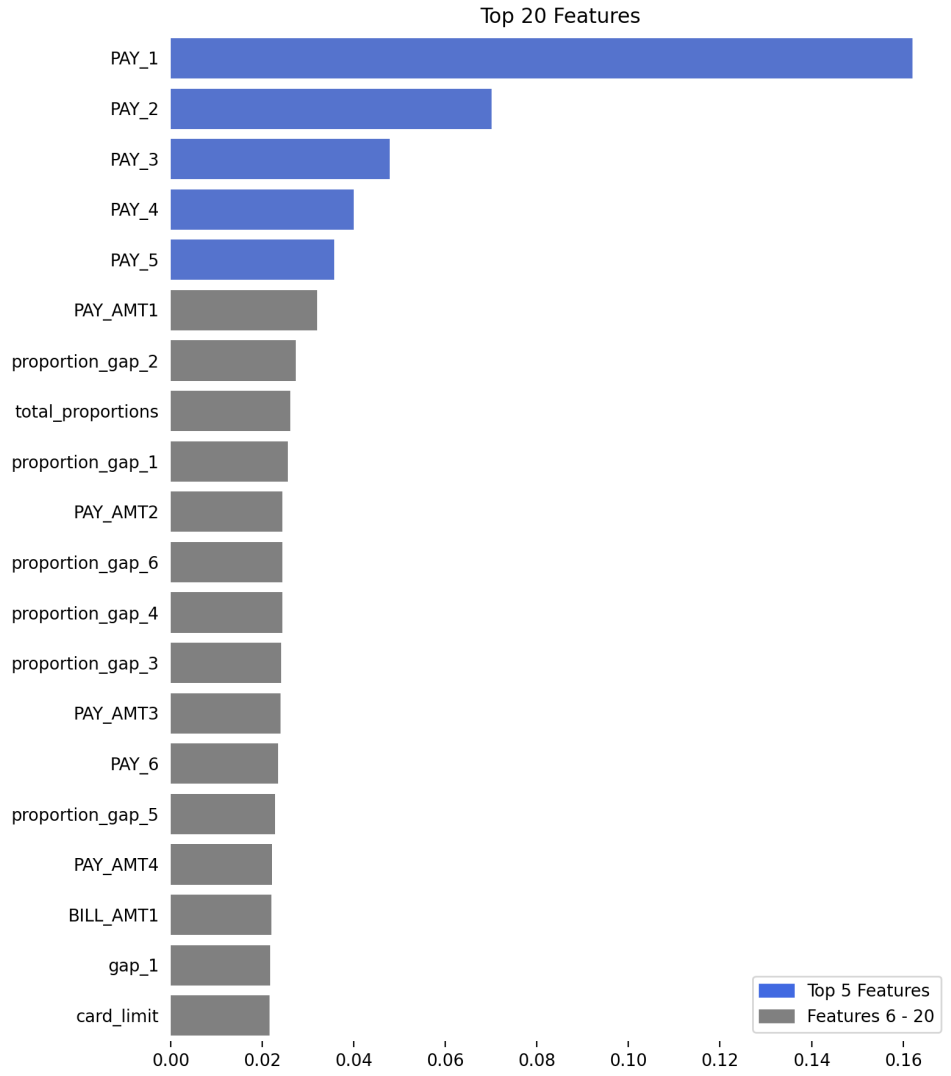
Part I – Random Forest

- Here is how the random forest ranks feature importance.



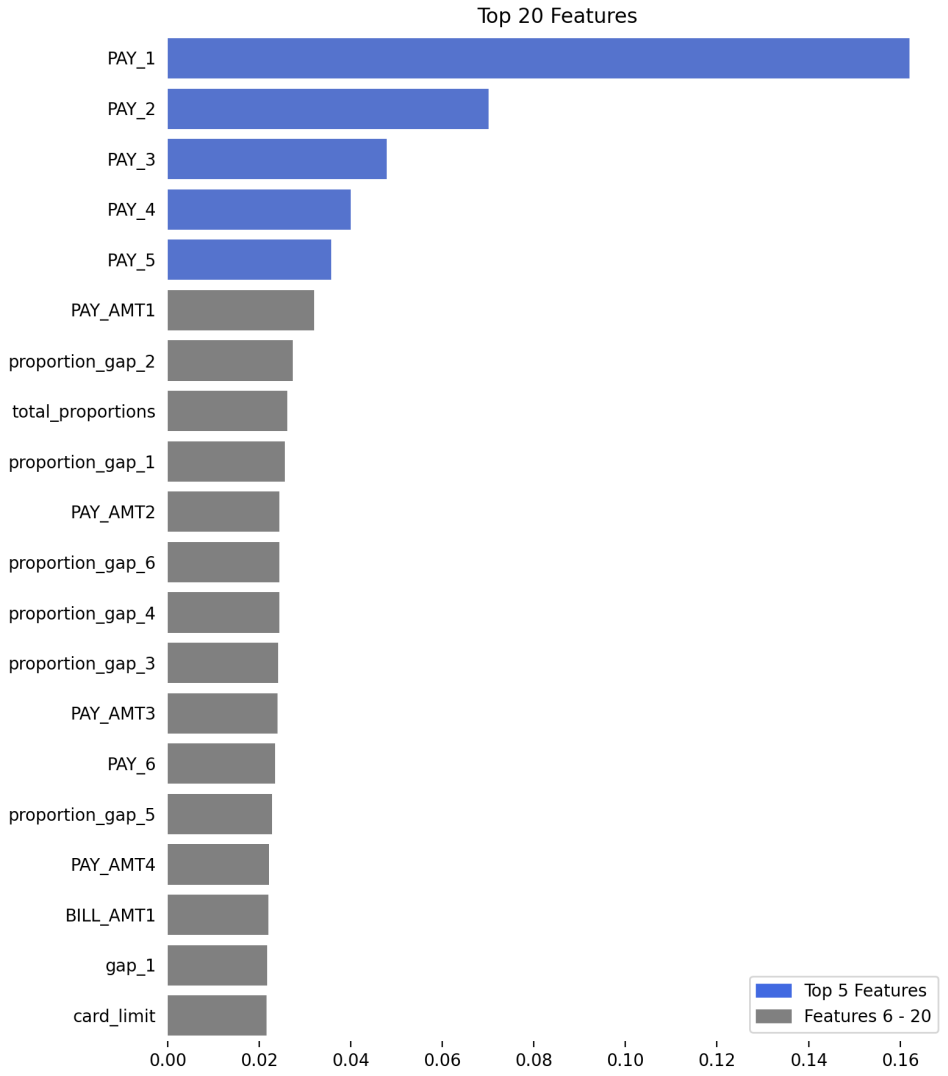
Part I – Random Forest

- PAY_1 – PAY_6 record how many months behind payment a customer was in a given month.
- PAY_1 is the most recent month.



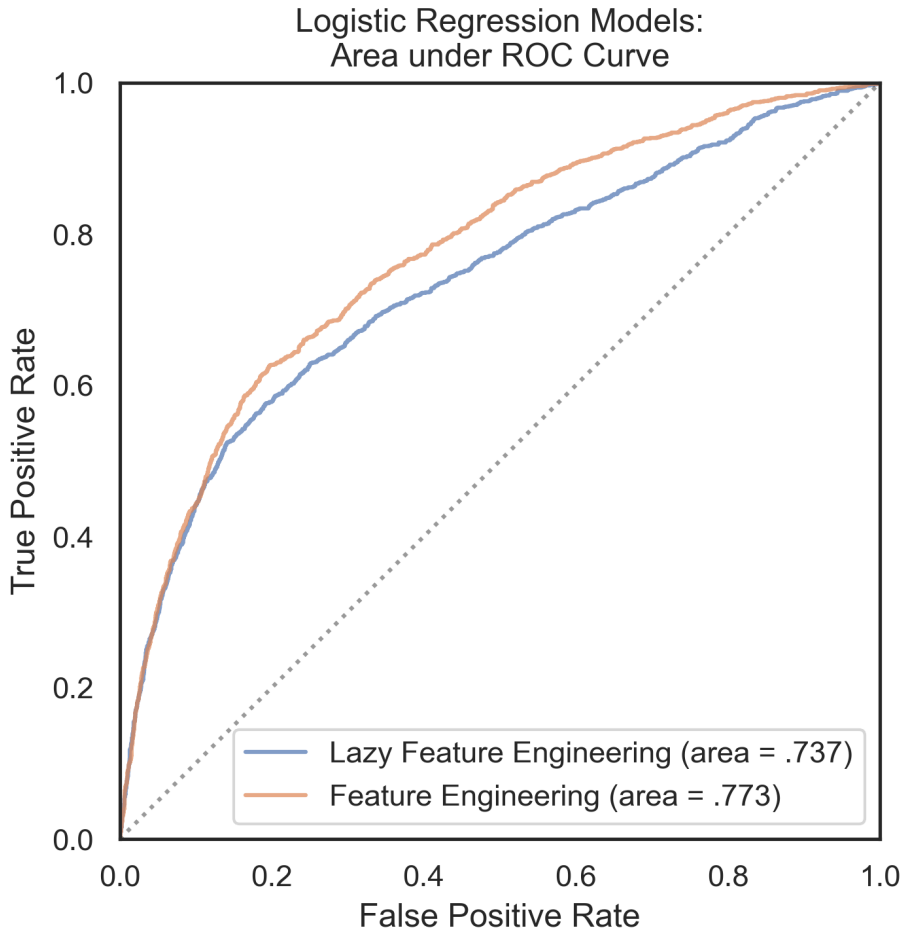
Part I – Random Forest

- Unsurprisingly, the best predictor of defaulting is missing payments.
- Payments missed recently matter more than payments missed several month ago.



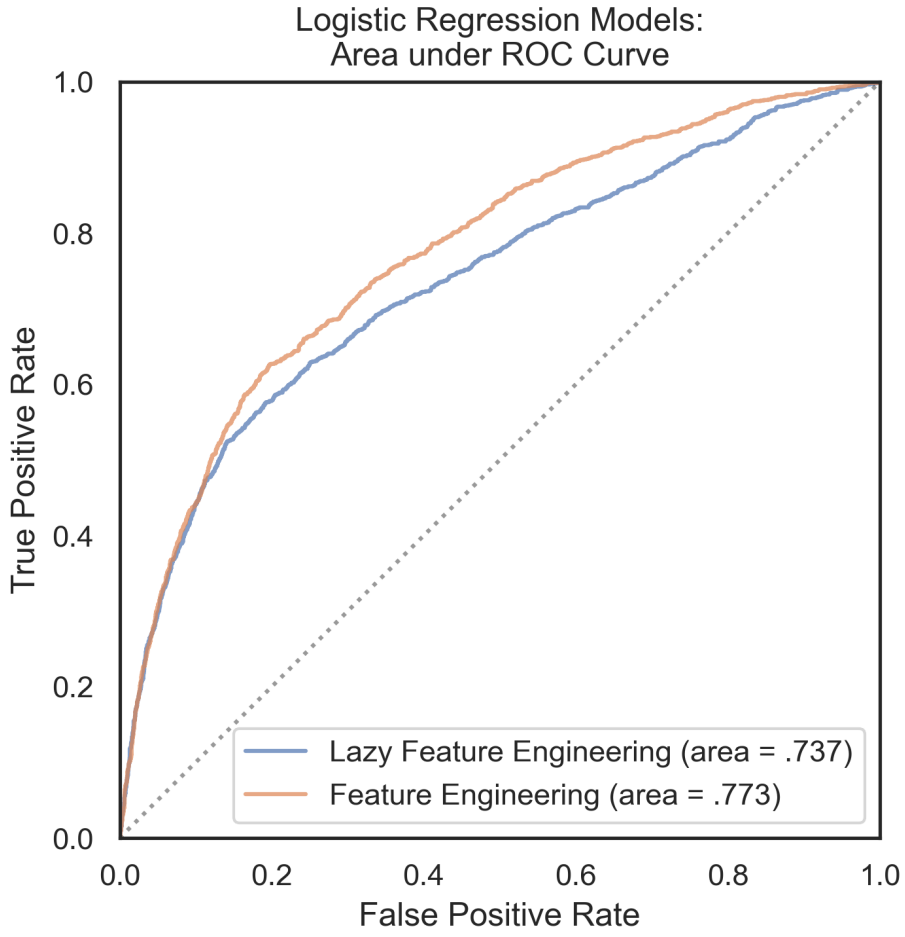
Part I – Logistic Regression (H)

- Unlike the random forest, the logistic regression model needed significant feature engineering.



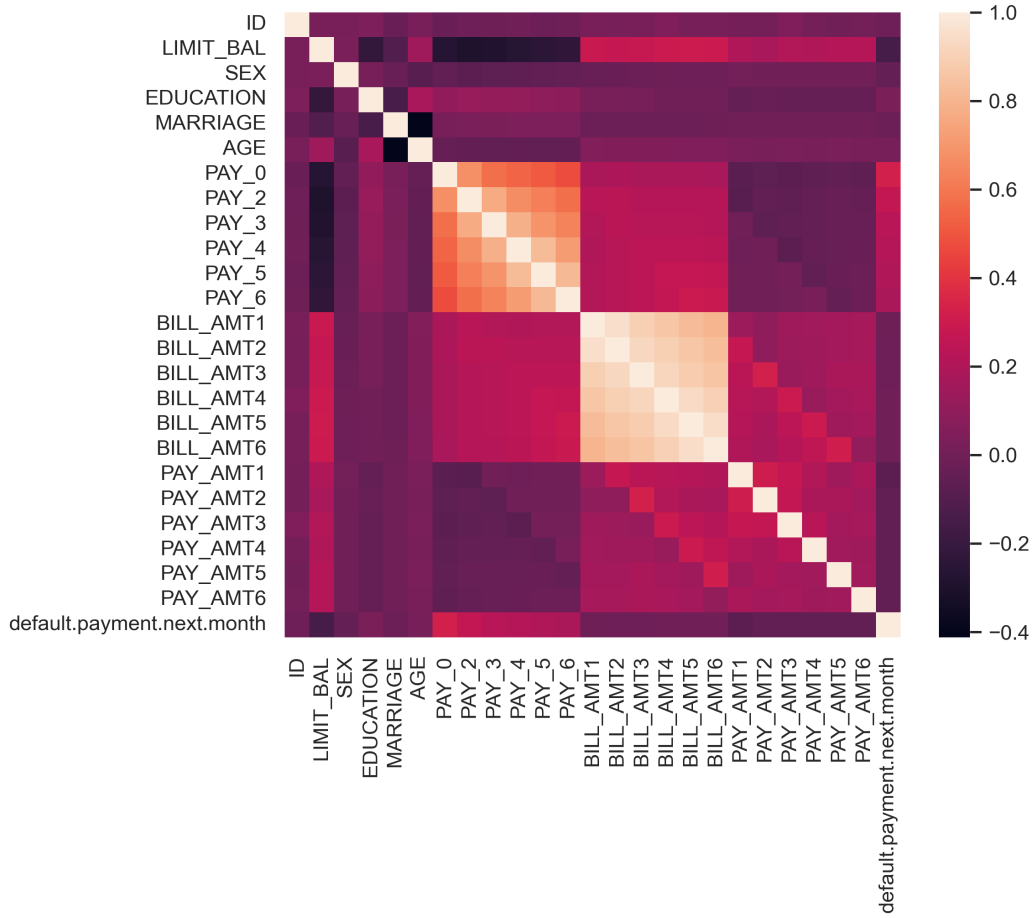
Part I – Logistic Regression (H)

- "Lazy Feature Engineering" meant only using the feature engineering that sklearn tools can perform.
- E.g., normalizing data, scaling it, and using PCA to transform it.
- It is lazy, because I did not have to think about the data.



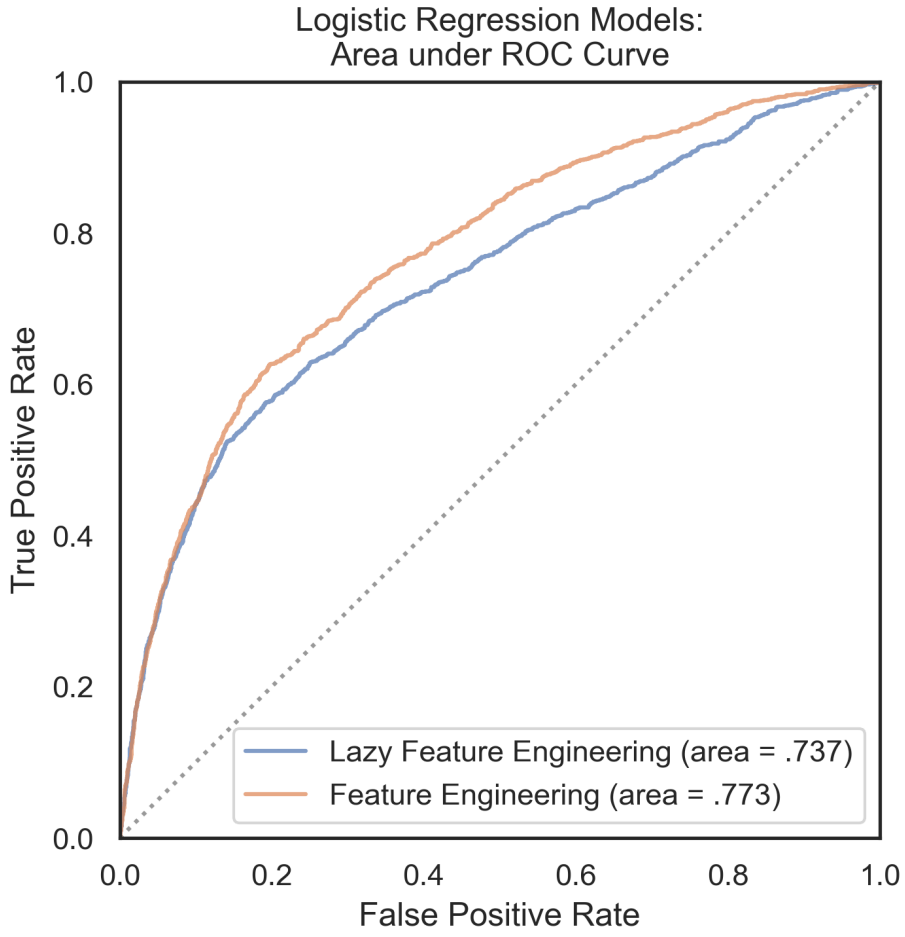
Part I – Logistic Regression (H)

- Some of this was necessary. For example, the data included several features with high collinearity.
- PCA helped eliminate some of that.



Part I – Logistic Regression (H)

- Model performance further improved after making sure that categorical features were not coded numerically.
- ROC AUC improved by .036.
- (E.g., using a coding such as “1: married; 2: single; 3: divorced.”)



Part I – Logistic Regression (H)

- Replacing this kind of numeric encoding was especially important, because some features mixed categorical and quantitative data.

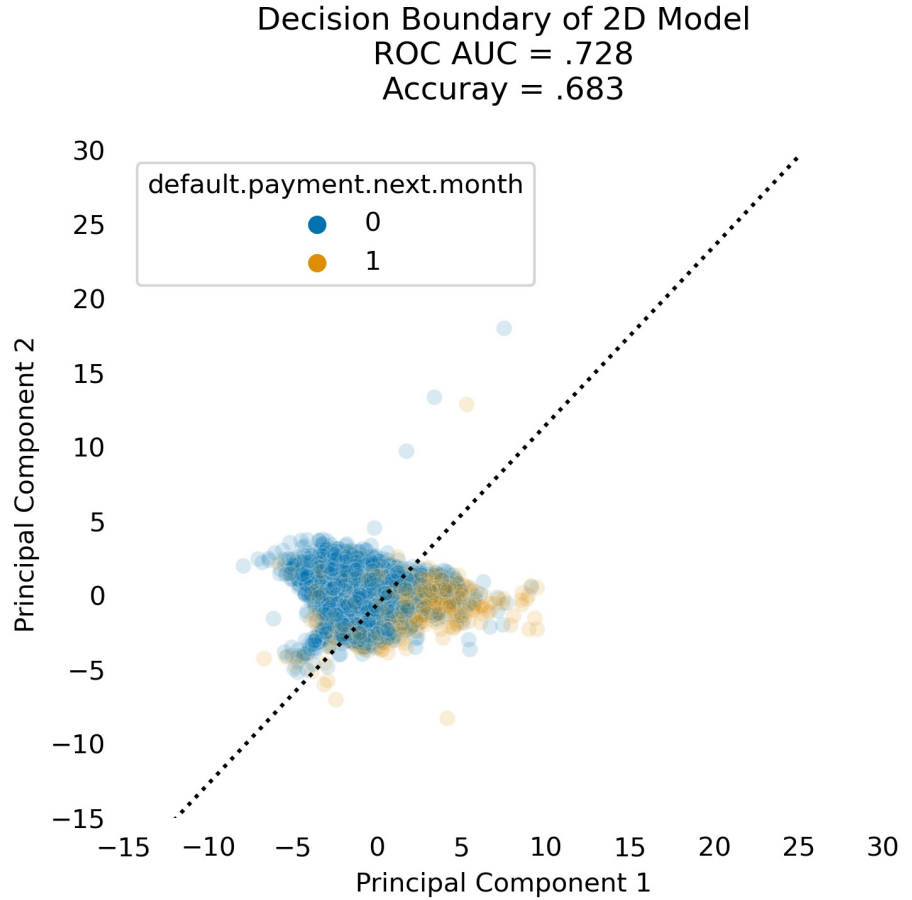
- Some columns would include positive numbers to indicate how many months late a customer was with payment.

- These very same columns would use negative numbers to indicate that a customer had, for example, fully paid their balance that month.

Part I – Logistic Regression (H)

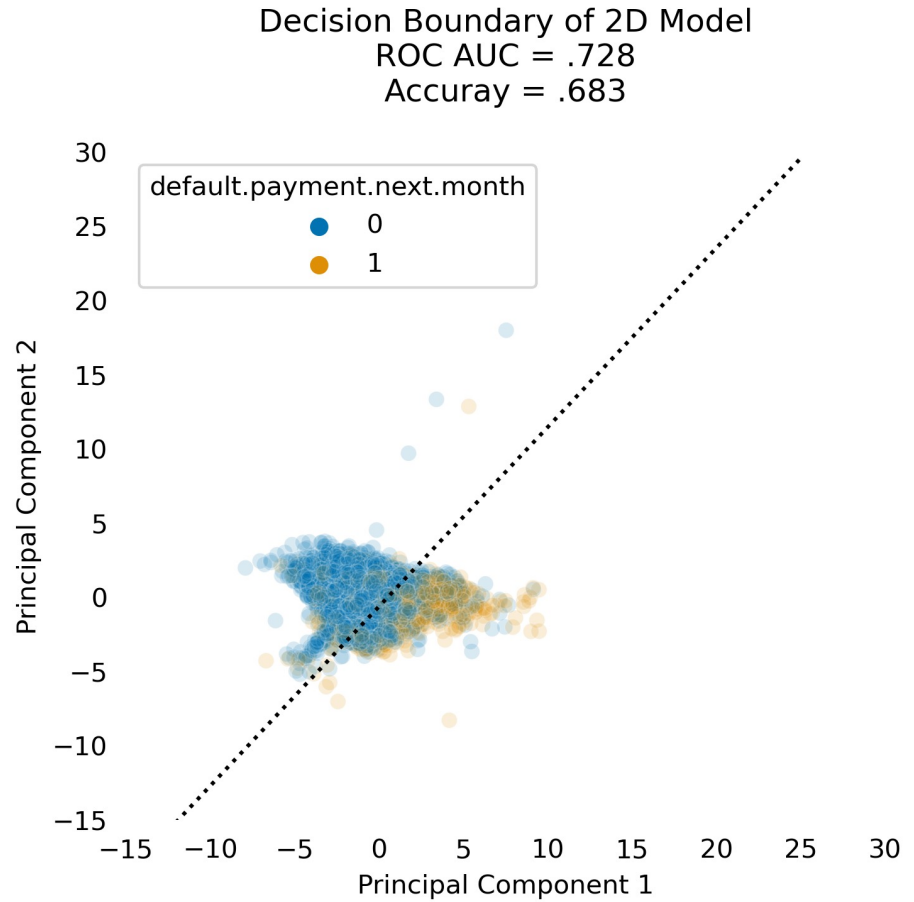
- It should also be noted that the logistic regression model achieved a higher ROC AUC by weighting the data to balance it.

- This means that model sacrificed accuracy for the sake on an improved ROC AUC.



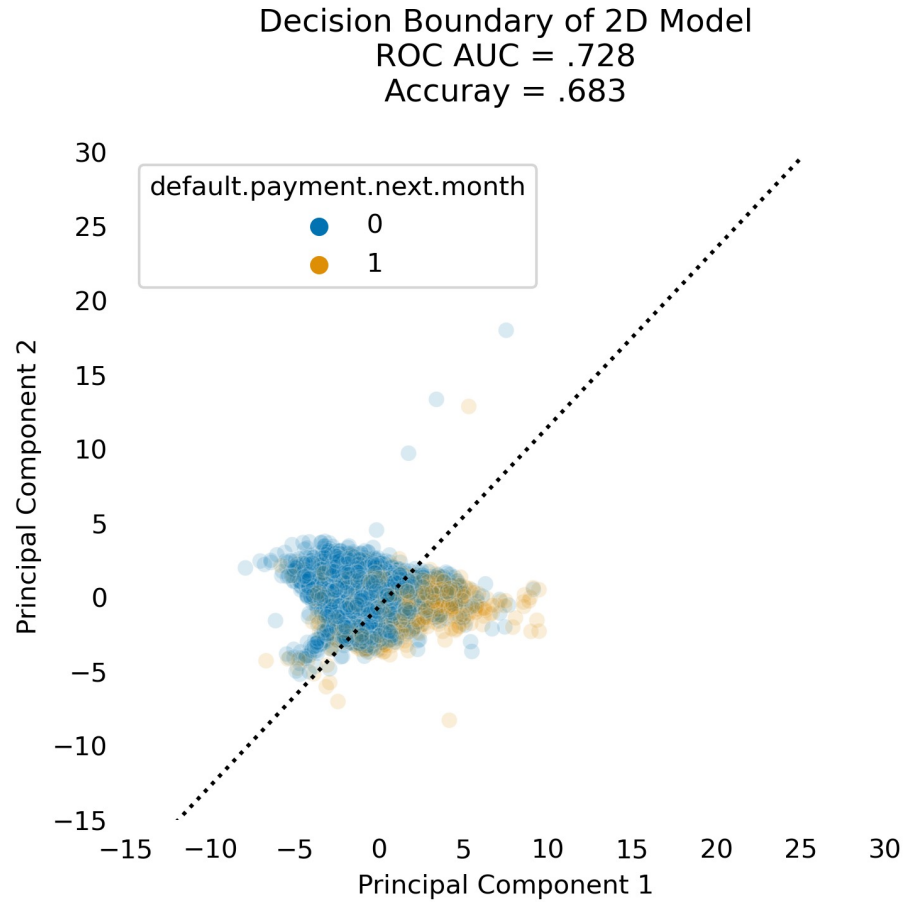
Part I – Logistic Regression (H)

- This 2D approximation of the Lazy Engineering Model gives a sense of how this works.
- Many non-defaulting points fall on the wrong side of the decision boundary.



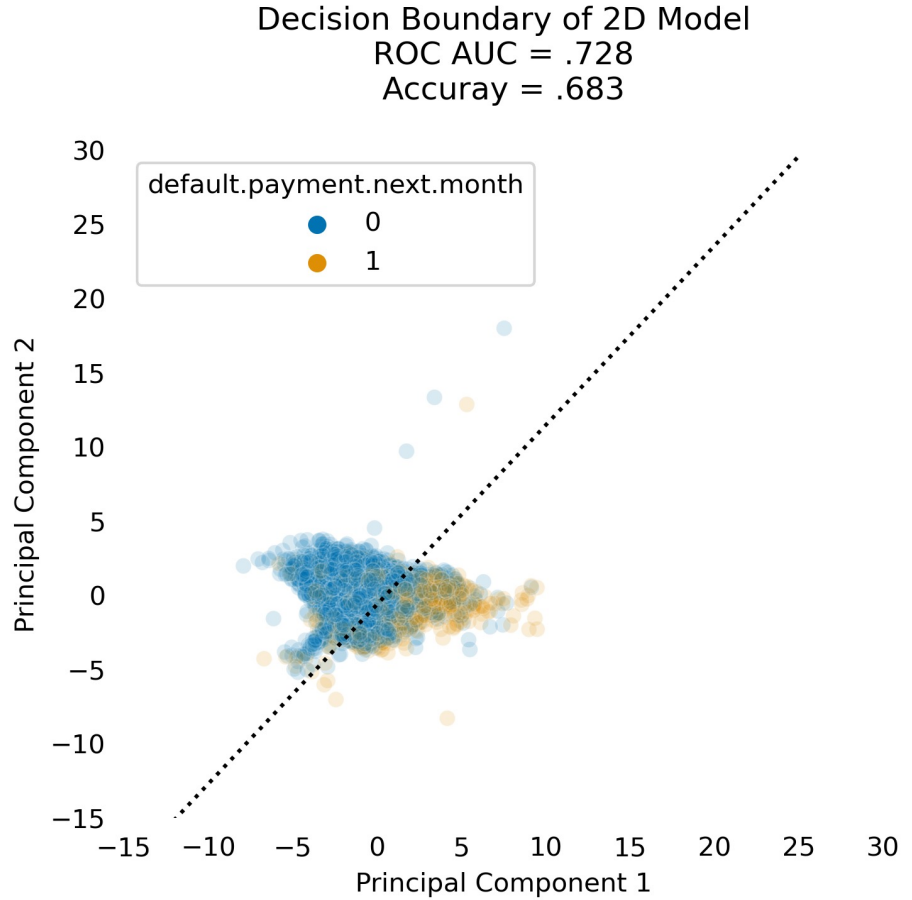
Part I – Logistic Regression (H)

- I explain why the 2D model is an acceptable approximation of the 21D model here (…).
- Though in part, this was an excuse to practice drawing decision boundaries.



Part I – Logistic Regression (H)

- Note that the final model did not sacrifice accuracy as substantially (accuracy = .751).
- It also uses 41 dimensions, so no pictures.



Part I – Logistic Regression (H)

- The logistic regression model largely agrees with the forest about what the most important features are.

```
[('PAY_1', 0.1969602968823826),
 ('PAY_2', 0.1951322177808584),
 ('PAY_4', 0.1881519317948492),
 ('PAY_3', 0.1874012852716690 2),
 ('PAY_5', 0.180423276211641 4),
 ('PAY_6', 0.1613086537575573 2),
 ('total_proportions', 0.136068739850460 75),
 ('proportion_gap_3', 0.122749898564862 06)]
```

```
[('PAY_5', 0.2822516105656802 7),
 ('PAY_4', 0.2806733146049443),
 ('PAY_2', 0.2798371065401702),
 ('PAY_3', 0.2794860746852618),
 ('PAY_1', 0.2688262045929791),
 ('PAY_6', 0.2538705656562868 5),
 ('married', 0.2404449761629470 8),
 ('card_limit', 0.2287237193102573 4)]
```

Part I – Logistic Regression (H)

- These are the top 8 features for Principal Components 1 and 3.
- These are two components that the model weights most heavily.
- (Coefficients = .302 and .295).

```
[('PAY_1', 0.1969602968823826),
 ('PAY_2', 0.1951322177808584),
 ('PAY_4', 0.1881519317948492),
 ('PAY_3', 0.1874012852716690 2),
 ('PAY_5', 0.180423276211641 4),
 ('PAY_6', 0.1613086537575573 2),
 ('total_proportions', 0.136068739850460 75),
 ('proportion_gap_3', 0.122749898564862 06)]

[('PAY_5', 0.2822516105656802 7),
 ('PAY_4', 0.2806733146049443),
 ('PAY_2', 0.2798371065401702),
 ('PAY_3', 0.2794860746852618),
 ('PAY_1', 0.2688262045929791),
 ('PAY_6', 0.2538705656562868 5),
 ('married', 0.2404449761629470 8),
 ('card_limit', 0.2287237193102573 4)]
```

Part I – Logistic Regression (I)

- The more interpretable logistic regression model has the worst performance in terms of ROC AUC.
- It performs roughly as well as the Random Forest in terms of pure predictive accuracy, however.
- The model was trained for accuracy rather than ROC score. We want to be able to make conclusions about odds of defaulting from this model.

Part I – Logistic Regression (I)

- The first interpretable model I trained had higher scores.

- Accuracy = .821
- ROC AUC = .758

- However, there were questions about whether the desired interpretation would be justified.

Part I – Logistic Regression (I)

- The goal is to be able to say how much a change in a given feature would change the odds that a customer defaulted.

- E.g. “For each month that a customer is late in the most recent month, their odds of default increase by a factor of 2.5.”

Part I – Logistic Regression (I)

- “For each month that a customer is late in the most recent month, their odds of default increase by a factor of 2.5.”

- These kinds of claims rely on a linear relationship between the feature and the log-odds of the target.

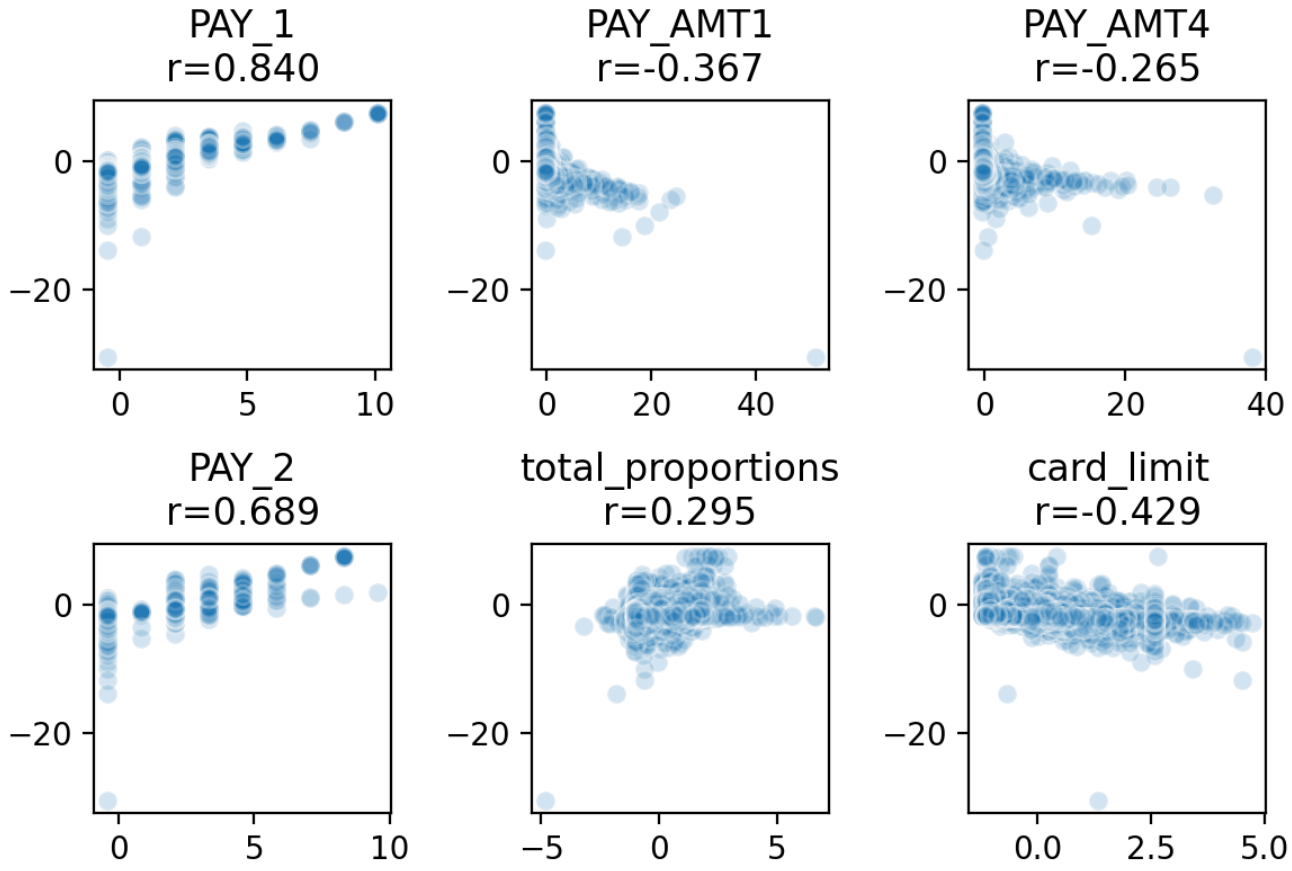
Part I – Logistic Regression (I)

- Such a linear relationship between features and log-odds is in fact an assumption of linear regression, but that assumption is arguably less important if we are only worried about predictive performance, and not what the model “means.”

- In this case, however, we are making claims based on the supposed meaning of the coefficients of our model.

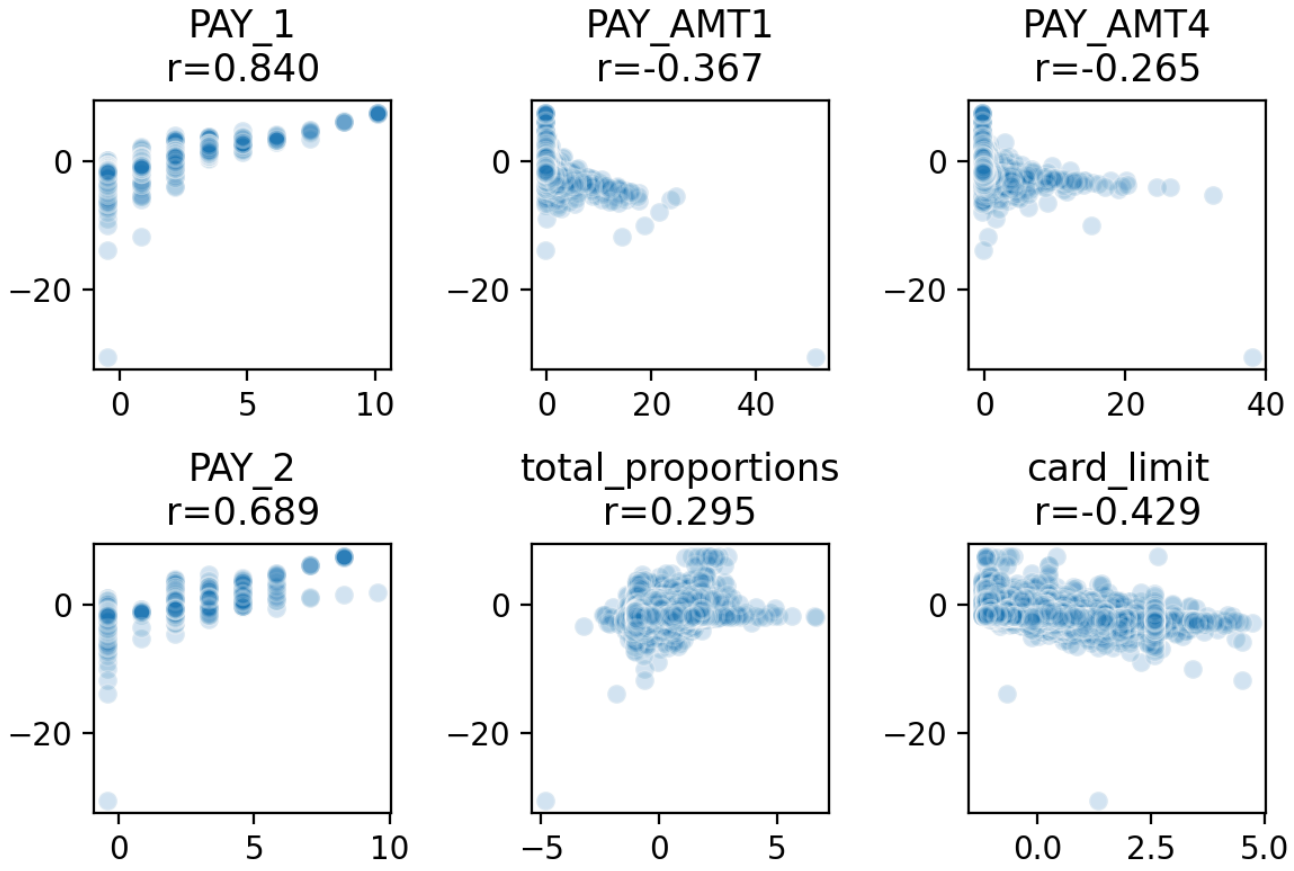
Part I – Logistic Regression (I)

- The first "interpretable" model I trained failed to meet the assumption of a linear relationship between the features and the log-odds.



Part I – Logistic Regression (I)

- Looking at 6 of the 15 features, we see that several do not have a linear relationship to the log odds.



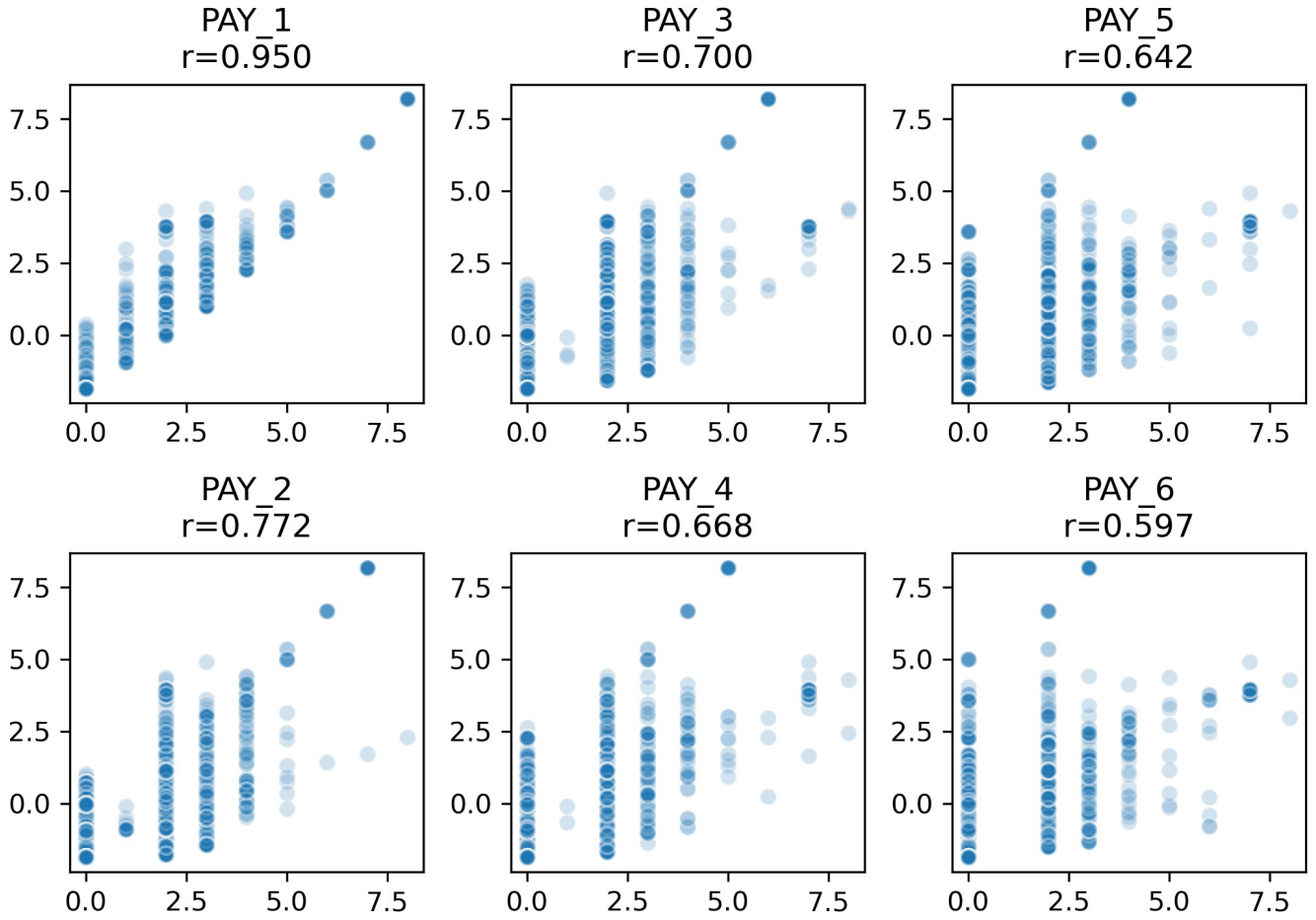
Part I – Logistic Regression (I)

- This led to a second model, which used only 6 features, all of the same sort.

- The model was based entirely on how many months behind their payments a customer had been in the previous six months.

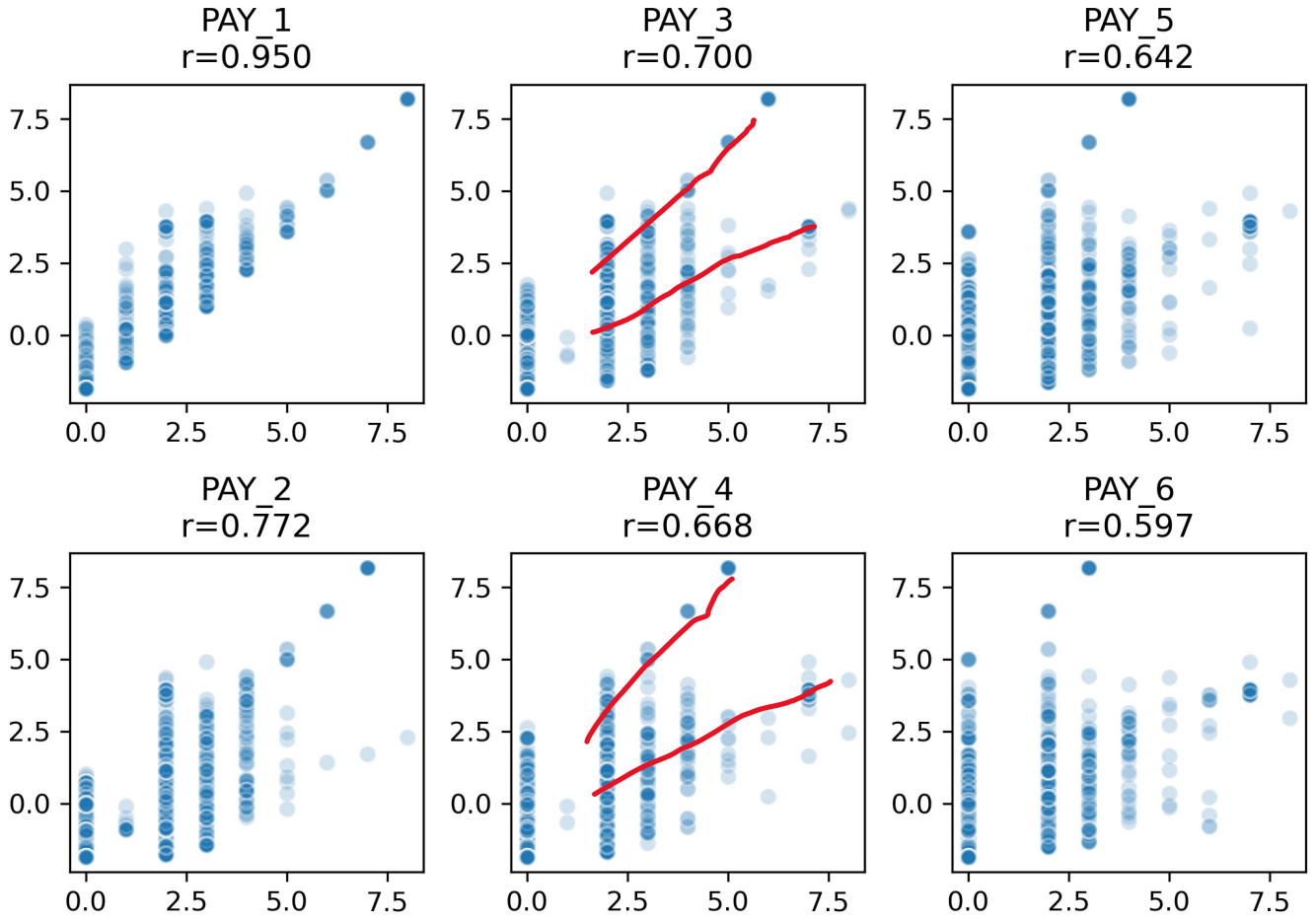
Part I – Logistic Regression (I)

- With this very sparse model, the correlation between the variable and the log odds is greater than .5 in each case.



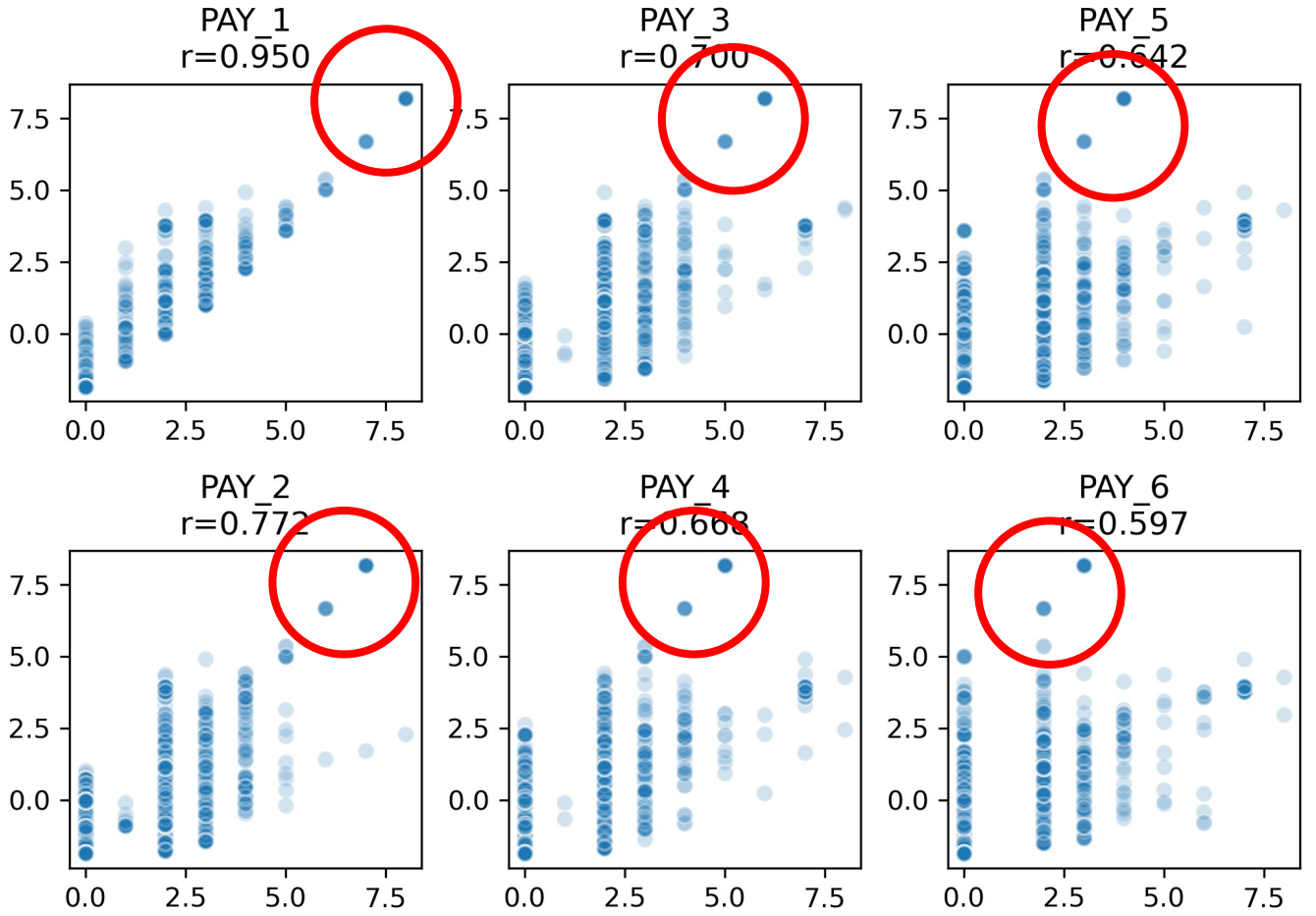
Part I – Logistic Regression (I)

- However, some of the features, such as PAY_3 and PAY_4 seem to have two trendlines.
- How big a problem is this?



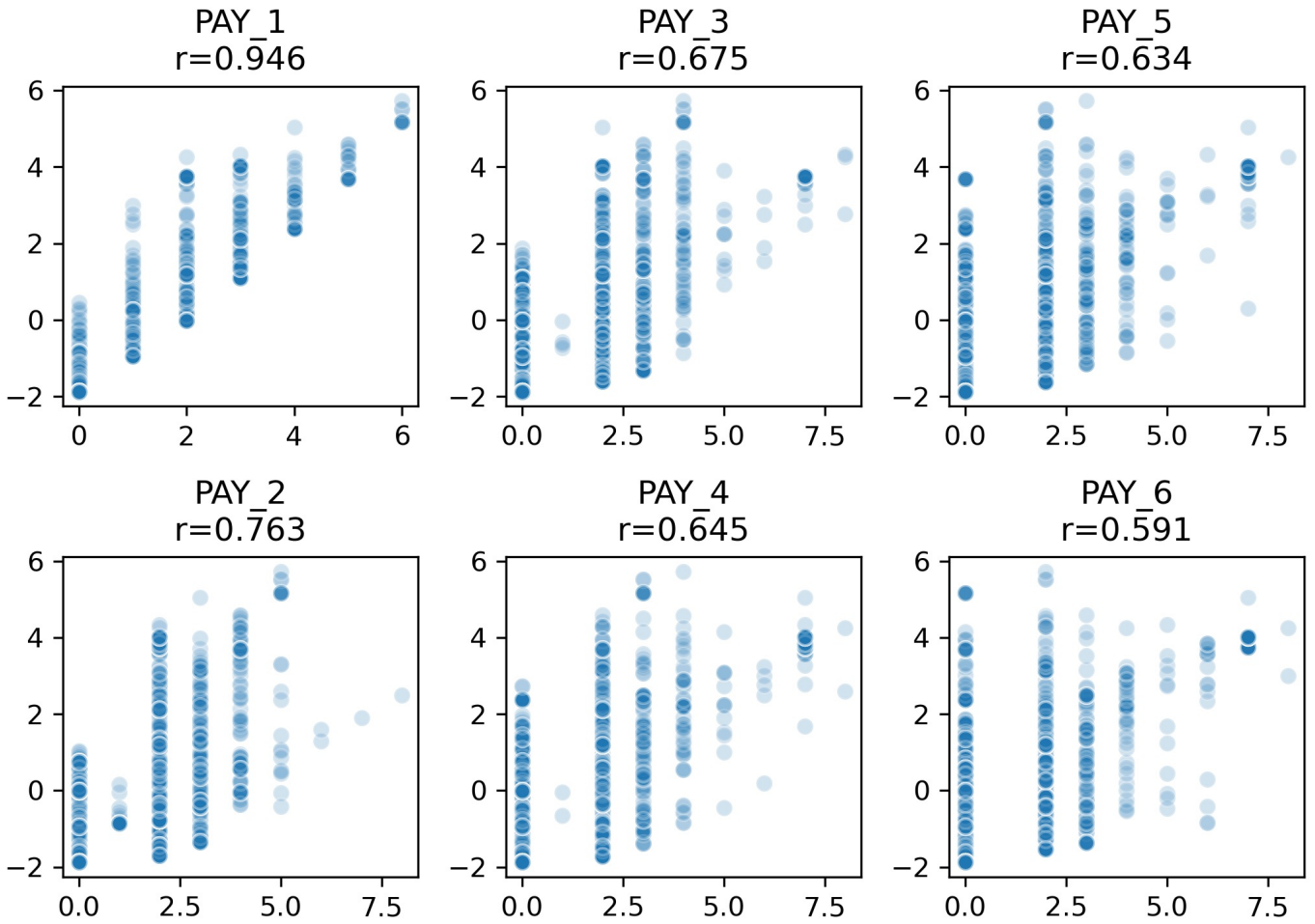
Part I – Logistic Regression (I)

- The circled dots represent the same customers.
- A customer who is 8 months behind in PAY_1 will be 7 months behind in PAY_2, and 6 months behind in PAY_3, etc.



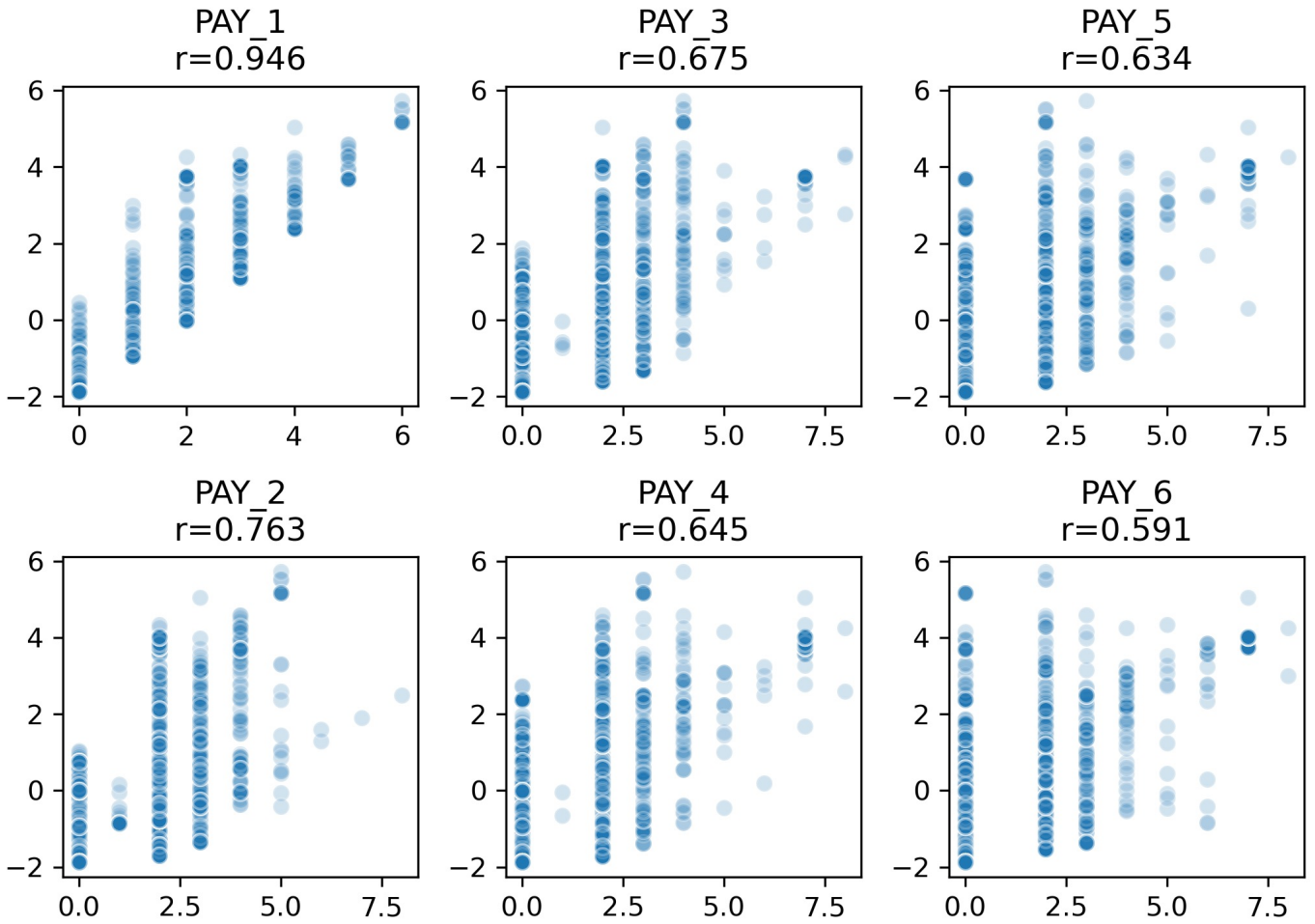
Part I – Logistic Regression (I)

- If we build a model that drops the datapoints, the appearance of a second trend in PAY_3 and PAY_4 disappears.



Part I – Logistic Regression (I)

- The coefficients remain fairly similar, even after dropping these datapoints.
- I am going to say that the assumption of linearity is met *well enough*.



Part I – Logistic Regression (I)

- Based on this model, we can say how being behind in payments affect the odds of default.

- Each month by which a customer is behind in the most recent month increases the odds of default by a factor of 2.5.

```
PAY_1            2.515569
PAY_2            1.049337
PAY_3            1.160510
PAY_4            1.098837
PAY_5            1.111922
PAY_6            1.195105
starting_odds    0.156220
dtype: float64
```

Part I – Logistic Regression (I)

- Each month by which an employee was behind 2 months ago increases the odds of default by a factor of approximately 1.05.

```
PAY_1              2.515569
PAY_2              1.049337
PAY_3              1.160510
PAY_4              1.098837
PAY_5              1.111922
PAY_6              1.195105
starting_odds      0.156220
dtype: float64
```

Part I – Summary of Test and Validation Scores

	ROC_AUC_test	ROC_AUC_valid	Accuracy_test	Accuracy_valid
Random Forest	0.788	0.782	0.819	0.821
Logistic Regression - (Heavy Processing)	0.772	0.772	0.753	0.751
Logistic Regression - (Easy to Read)	0.725	0.741	0.821	0.819

- This concludes Part I
- Once again, this is the performance of the three final models.

Part II – Tests and Simulations

- Part II involved statistical tests and simulations on the data.
- I tested a total of seven hypotheses (more on the problems that creates in a minute).
- In some cases the assumptions of the relevant tests were not met by the data, and so I wrote bootstrapping or permutation simulations to determine whether observed results were significant.

Part II – Inadvertent p-hacking?

- Testing 7 different hypotheses runs the risk of inadvertent p-hacking.
- If we set α = .05 and test 7 hypotheses, we have a 30% chance of getting at least one significant result by chance.
- To reduce the risks of inadvertent p-hacking, I am setting α very low, to .007.
- The chance of getting at least one significant result by chance is 4.8%.

Part II – Inadvertent p-hacking?

- Another concern is that I was going to decide which hypotheses to test after exploring the data.

- Again, to guard against p-hacking, I will initially explore a random subsample of 10,000 cases. This exploration will include tests and simulations.

Part II – Inadvertent p-hacking?

- On the basis of these exploratory tests and simulations, I will decide which tests and simulations to run on the remaining 20,000 cases.

- These tests will then provide confirmation, or not, of the results of the original findings.

Part II – Main Question and Assumptions

- An issue briefly discussed in pp. 5-7 of this slideshow, is whether the data represents a random sample of customers.

- Approximately 22.2% of customers default in a six-month period.

- The median customers' spending over the period exceeds payments by an amount totaling to 1.4x their credit limit.

Part II – Main Question and Assumptions

- My working assumption is that the data represents customers who have been flagged as being at higher risk of default.

- If this assumption is true, there may be some concern that women make up around 60% of the sample population.

Part II – Main Question and Assumptions

- First, we may worry that women are more likely to be flagged than men.

- Furthermore, we may worry that women are more likely to be flagged even when the available data indicates that they are at less risk of defaulting.

- If the latter is true, it would suggest obvious ways to improve policy: either flag more men, or flag fewer women.

Part II – Investigating the Main Question

- Out of the initial subsample of 10,000 customers, 6042 are women.
- According to Wikipedia, 51.5% of Taiwanese are women.
- The z-score on this difference is 18.24.

Part II – Investigating the Main Question

- A test is admittedly unnecessary (z=18.24!). But I ran one to get practice with the software.

- A two-sided binomial test indicates that this difference is statistically significant.

- p = 7.5e-72

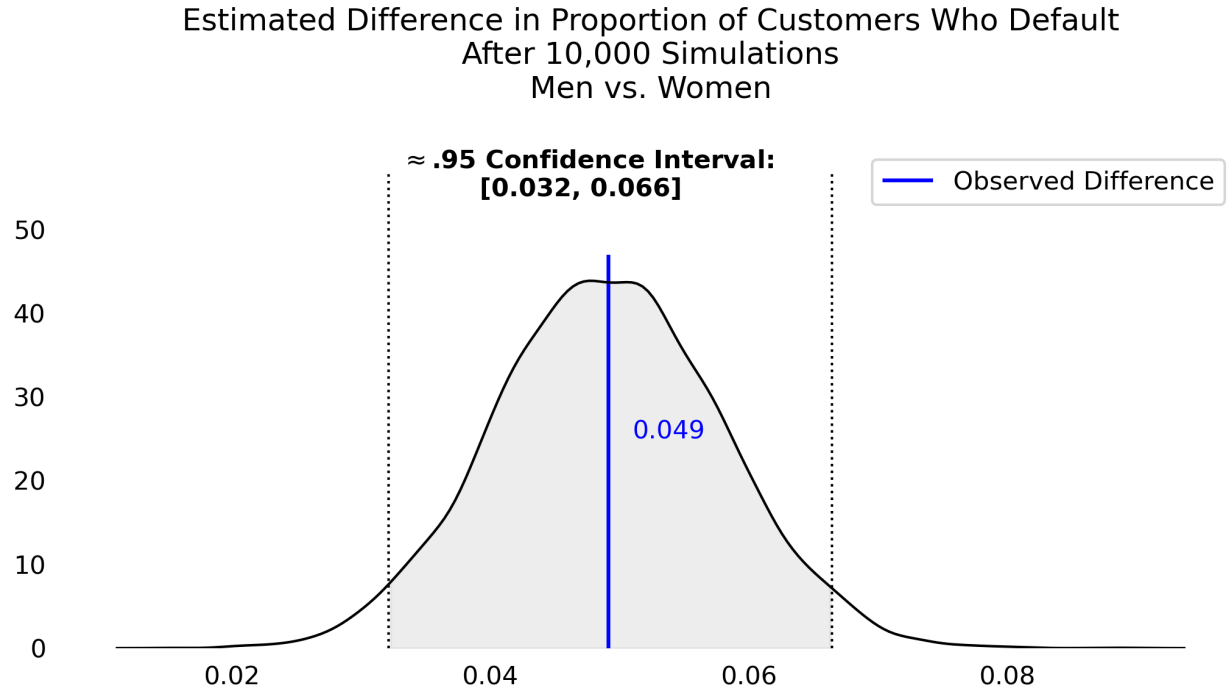
Part II – Investigating the Main Question

- A test is admittedly unnecessary (z=18.24!). But I ran one to get practice with using the software.
- A two-sided binomial test indicates that this difference is statistically significant.
- p = 7.5e-72. (This is a very small number.)
- According to our 95% confidence interval, women make up between 59.5% and 61.4% of the sampled population.

Part II – Investigating the Main Question

- Of our subsample, the default rate for men is .256.
- For women .207
- This is a difference of .049.

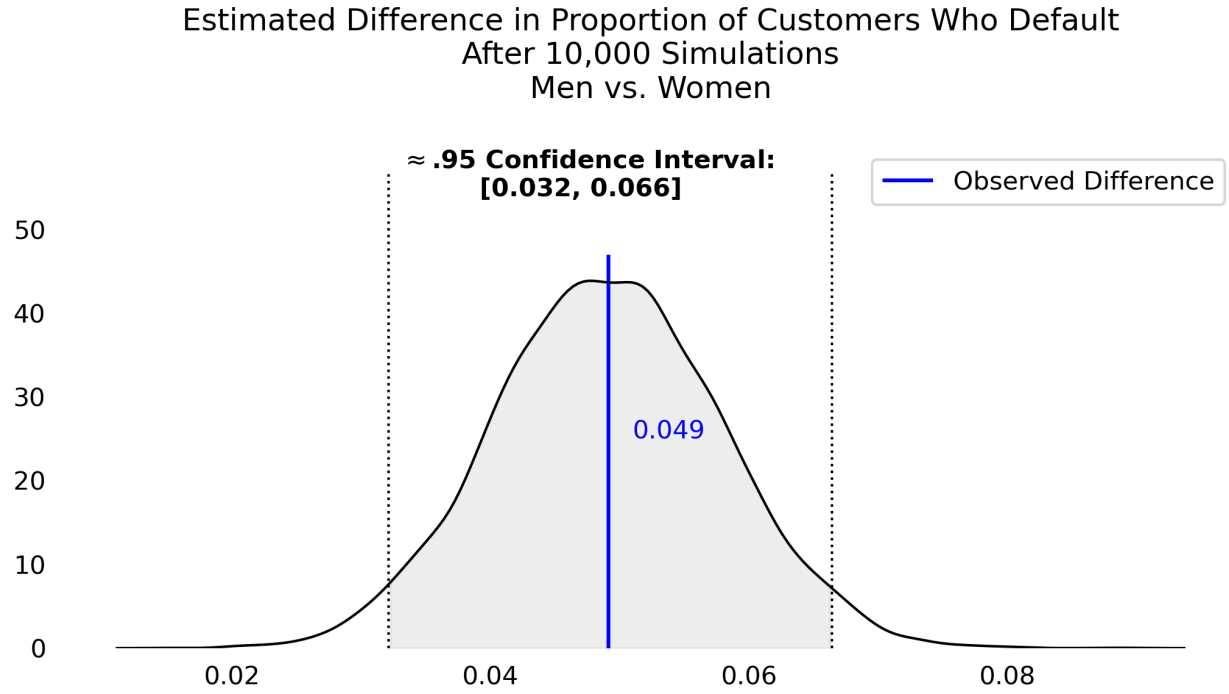
- A one-sided t-test confirms that this difference is statistically significant.
- p = 7.8e-9.



Part II – Investigating the Main Question

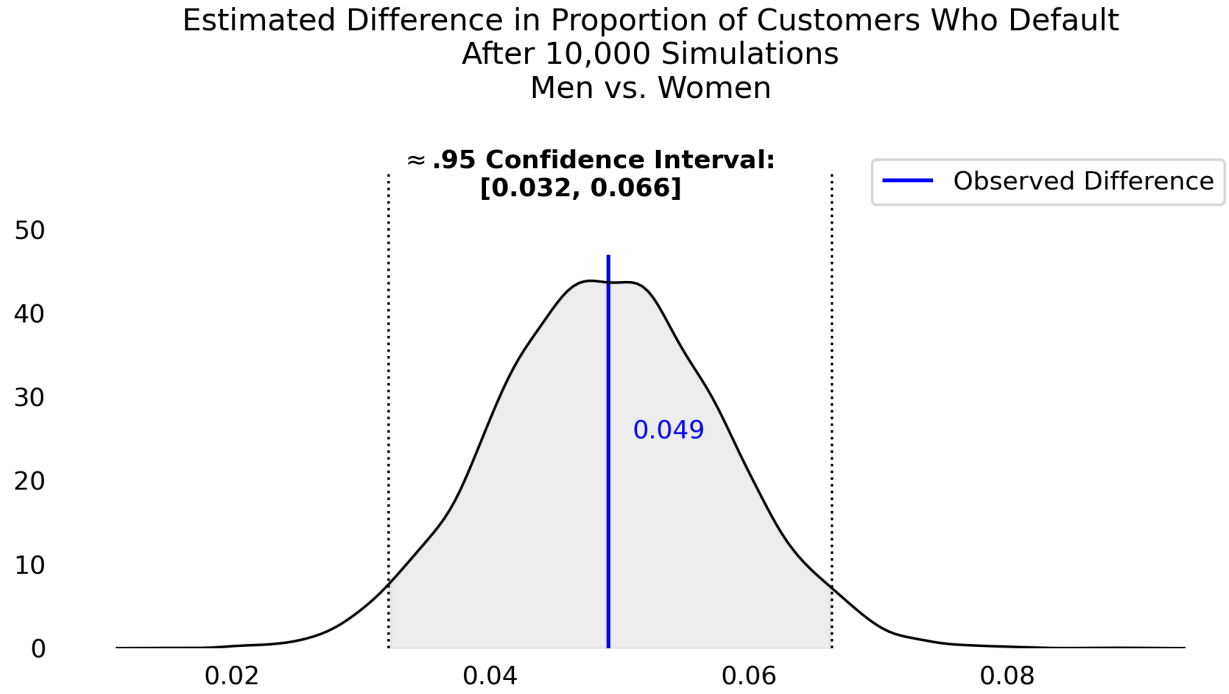
- One question is how to calculate the confidence interval.

- Using pooled variances, the .95 Confidence Interval of the difference in the proportions is [.041, .057].



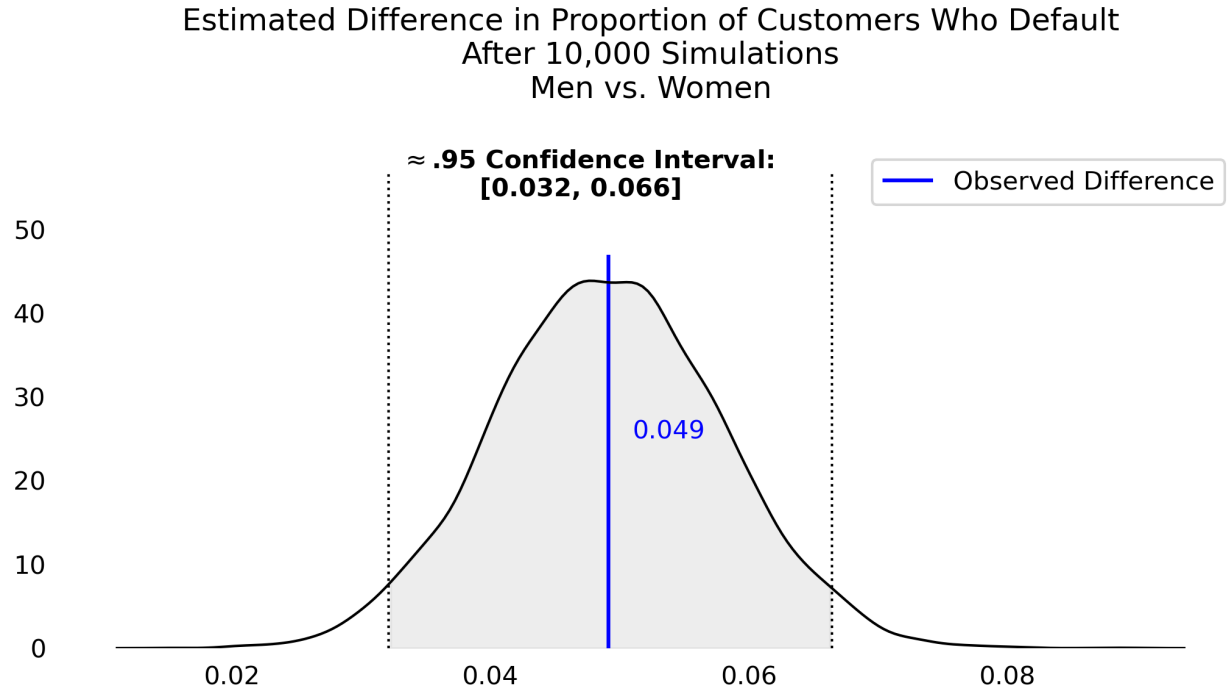
Part II – Investigating the Main Question

- However, one of the assumptions of pooled-variance is not met.
- Running 10,000 bootstrap simulations provides a more cautious estimate for the .95 CI (see figure).



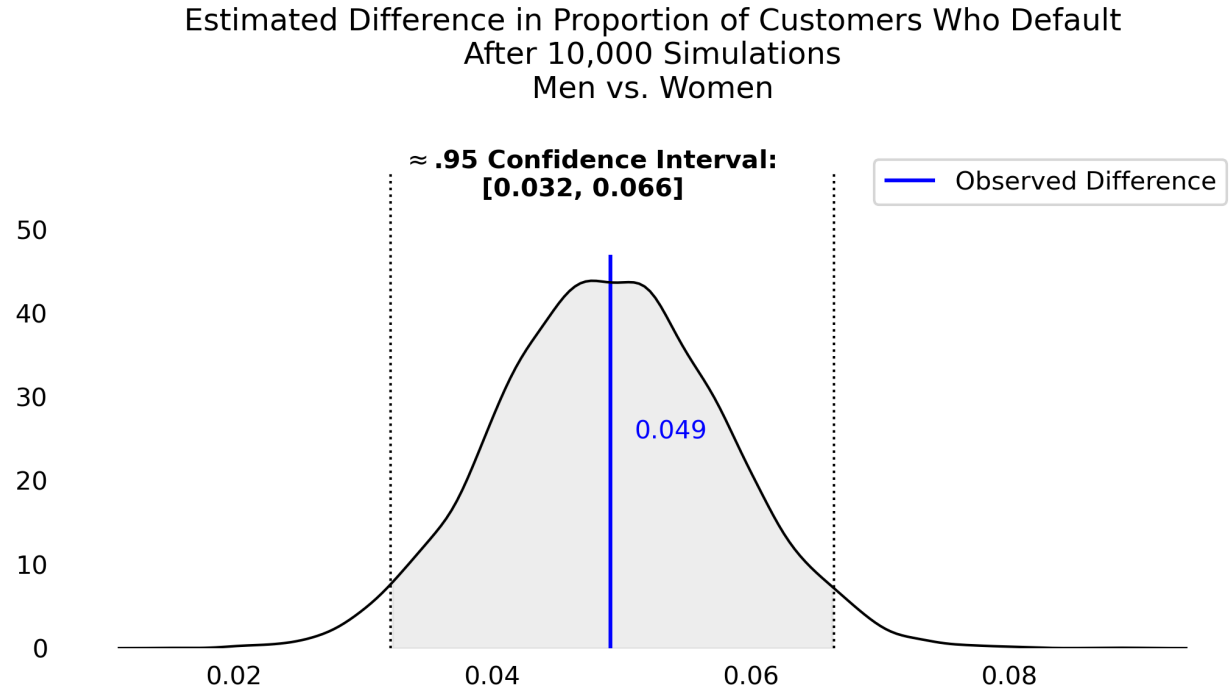
Part II – Investigating the Main Question

- This .95 CI is [.032, .066].
- This is the one I'll use for now.



Part II – Investigating the Main Question

- Assuming our working hypothesis, this is initial evidence that women are flagged at rates that do not reflect their risk of default.



Part II – Investigating the Main Question

- We could put the point this way, if we removed women who did not default from our subsample, until there was parity in the default rates, we would have reduced the number of women in the sample by 19.2%

- Women would make up 55.2% of the sample population, not 60.4%.

- Assuming our working hypothesis, the system for flagging at risk customers may overestimate the risk that women default—or underestimate the risk for men.

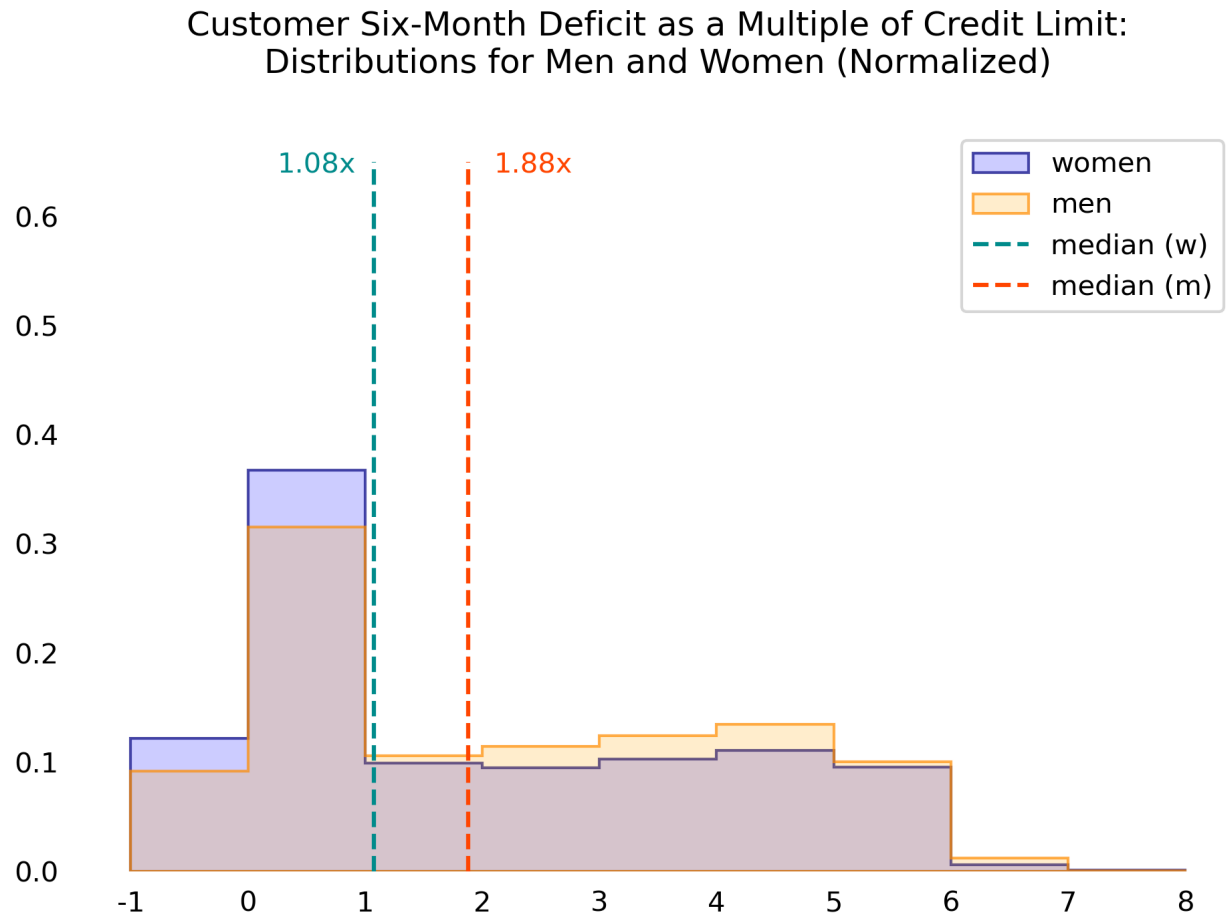
Part II – Investigating the Main Question

- I will consider some possible explanations.
- These explain some, but not all, of the discrepancy in the rates at which women appear to be flagged.
- Men from the sampled population are more likely to have exceeded their limit, to have exceeded it by more, and they are more likely to have missed a payment.

Part II – Investigating the Main Question

- Perhaps women are more likely to have exceeding their credit limit.

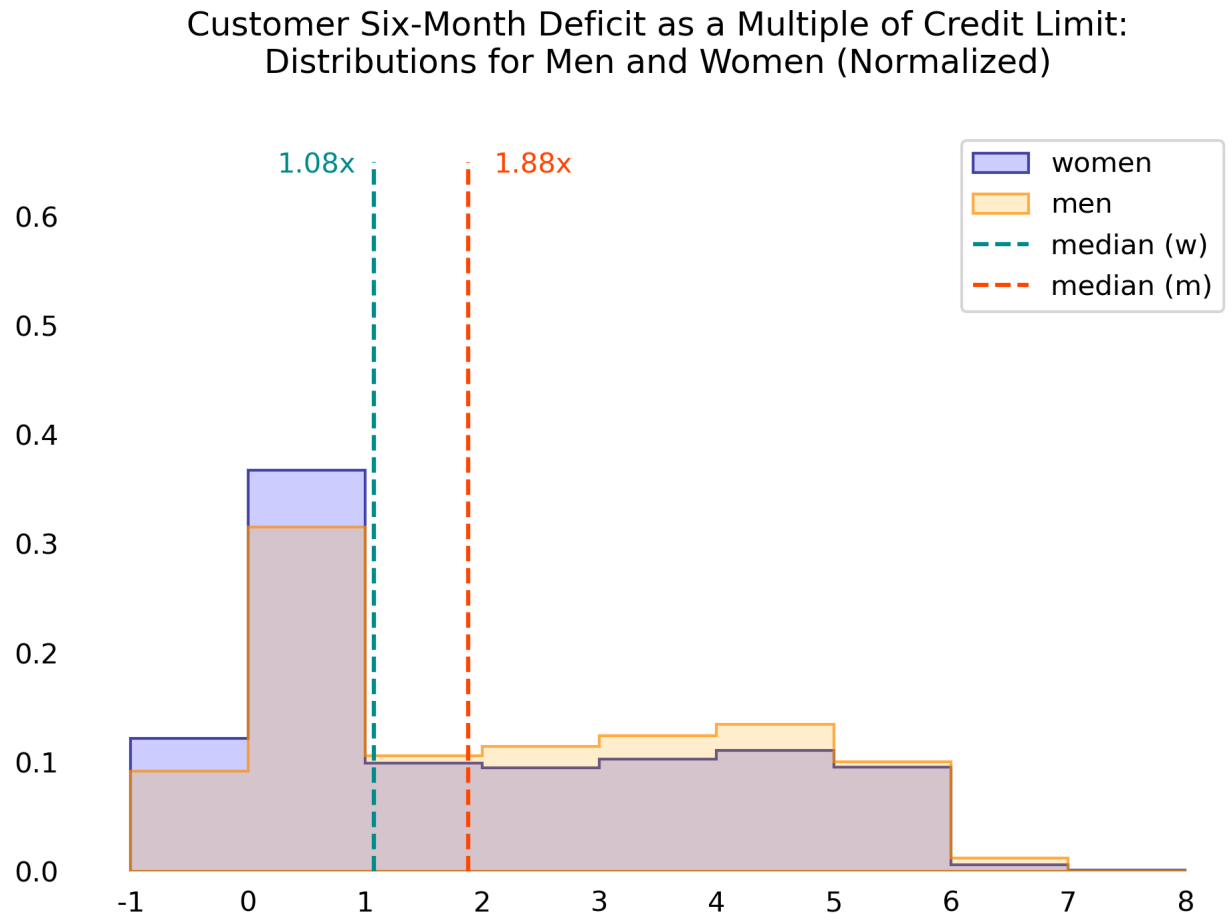
- But in the six-month period for which we have data, men are more likely to exceed their credit limit, and they do so by larger amounts.



Part II – Investigating the Main Question

- Note, by the way, that both the median woman and median man in the subsample have exceeded their credit limit.

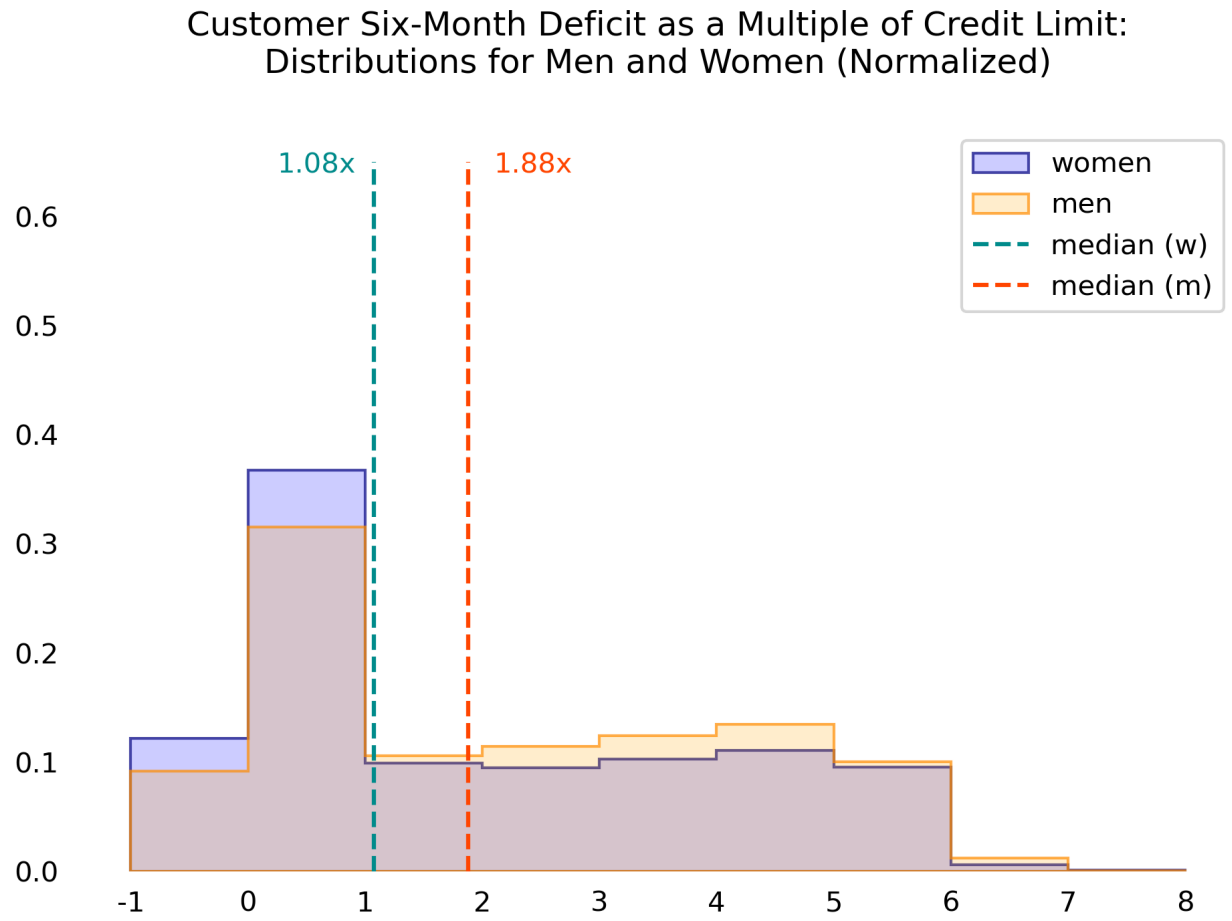
- This is why I suspect this sample represents higher-risk customers.



Part II – Investigating the Main Question

- Using Moody's median test from the scipy.stats library confirms that the difference in medians is statistically significant.

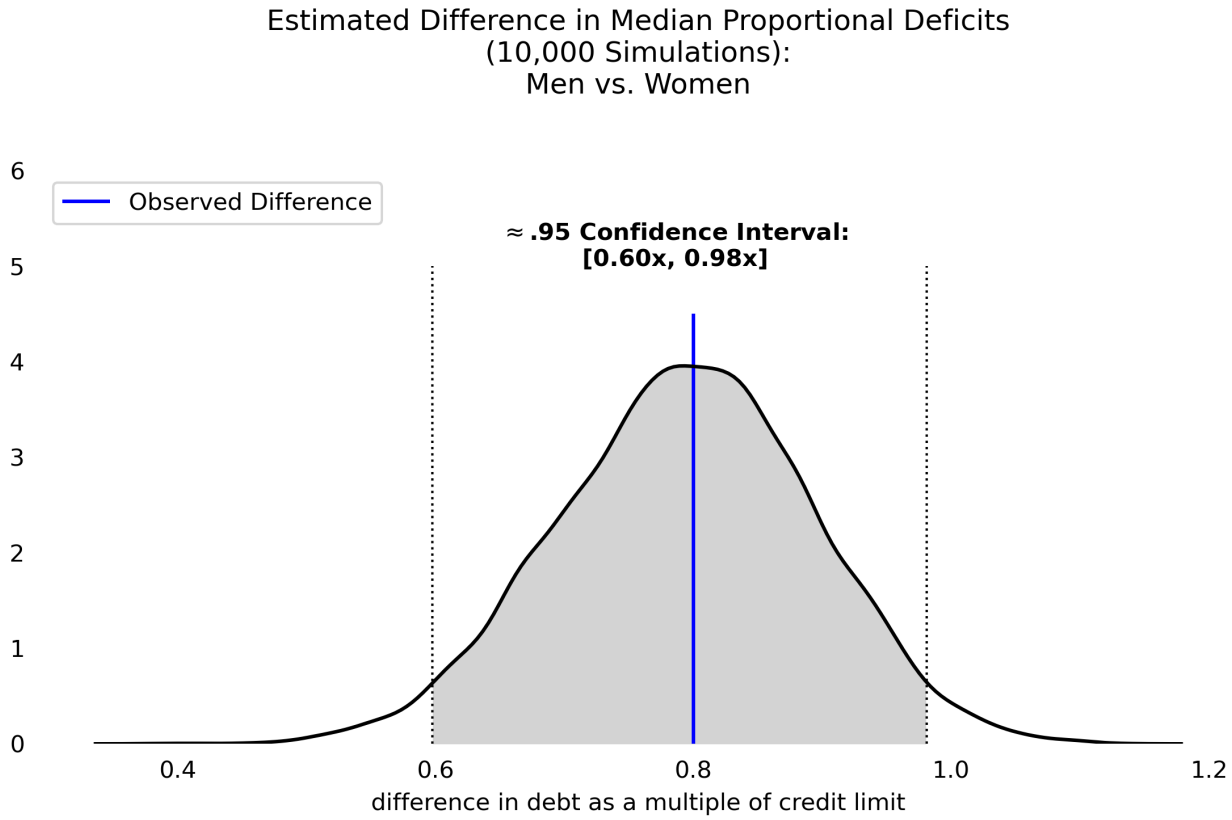
- p=1.7e-14



Part II – Investigating the Main Question

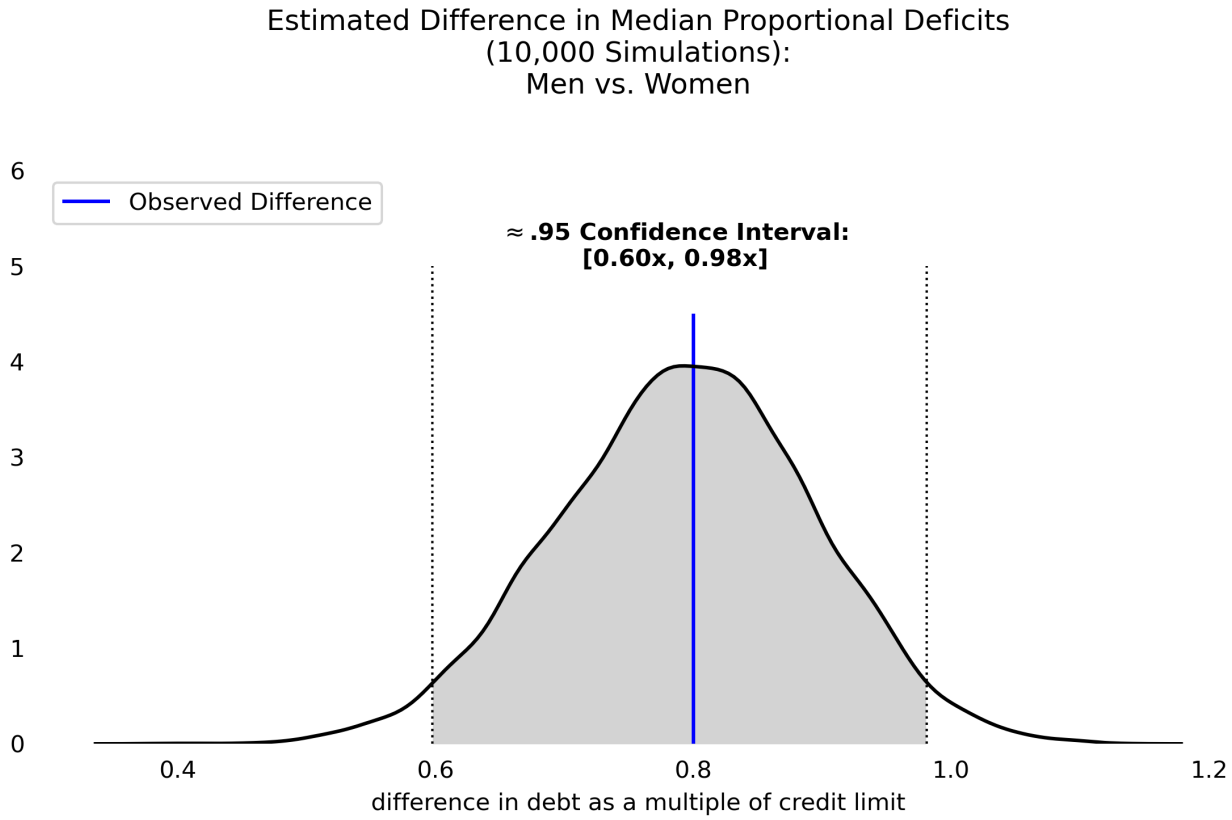
- I don’t know how to calculate confidence intervals for the difference in medians.

- I ran a bootstrap simulation to estimate a confidence interval.



Part II – Investigating the Main Question

- The .95 CI ranges from 0.60 to 0.98.
- In other words, the median woman from the sampled population would need additional debt equal to 60% to 98% of her credit limit to have proportional debt equal to that of the median man.



Part II – Investigating the Main Question

- Another possibility is that women are more likely to have exceeded their credit limit **or** missed a payment than men.

flagged	False	True
sex		
female	2174	3868
male	1122	2836

flagged	False	True
sex		
female	0.359815	0.640185
male	0.283477	0.716523

Part II – Investigating the Main Question

- But men in the subsample are more likely than women to have missed a payment or exceeded their credit limit.
- 71.7% vs. 64.0%

flagged	False	True
sex		
female	2174	3868
male	1122	2836

flagged	False	True
sex		
female	0.359815	0.640185
male	0.283477	0.716523

Part II – Investigating the Main Question

- A chi-square test confirms that this difference is statistically significant.

- p = 2.4e-15.

flagged	False	True
sex		
female	2174	3868
male	1122	2836

flagged	False	True
sex		
female	0.359815	0.640185
male	0.283477	0.716523

Part II – Investigating the Main Question

- Men are more likely to fall into at least one of the high-risk categories than women are.

- But the magnitude of the proportional difference is smaller than the difference in rates of default.

- So, by looking at a disjunction of flagged categories, we may explain **some of** the discrepancy in the proportion of men and women (again assuming that this is a sample of customers flagged as risky).

Part II – Investigating the Main Question

- As noted above, for the men and women in our samples to have the same rate of default, the number of women would need to go down by about 19.2%, all from the category of non-defaultees.

- For the proportion of sampled men and women in a high-risk category to be equal, the number of women in our sample would need to be 10.7% lower.

Part II – Investigating the Main Question

- This is still a noticeable difference, even if it is a smaller one.

- Again, on the assumption that the data represents a sample of high-risk customers, the company may wish to revise how it identifies these customers.

Part II – Verifying Our Conclusion

- Running the tests again on the larger subsample of 20,000 largely confirmed the earlier results.
- There is one notable exception, however.
- While men in this larger subsample are still more likely to default than women, the difference is notably smaller than in the previous subsample.

Part II – Verifying Our Conclusion

- 23.4% of men defaulted, compared to 20.8% of women.
- This is still a statistically significant difference.
- But a difference of .0026 in the proportions is outside of the .95 Confidence Interval calculated on the basis of the smaller subsample.

Part II – Verifying Our Conclusion

- It is well outside of the .95 CI calculated on the assumption of pooled variance. (On that assumption, results like this should occur in one out of 70 million cases.)

- It is even outside of the more conservative CI based on simulations, [.032, .066].

- Results of this extreme happened in (slightly) less than 1% of simulations.

Part II – Verifying Our Conclusion

- Other tests confirmed the earlier findings.

- Again, we can say that there is a discrepancy in the rates at which men and women default, exceed their credit limits, and miss payments, and the rate at which they are flagged as credit risks. (Again, assuming this is a sample of flagged customers.)

- But the discrepancy is a bit smaller in this subsample.

Part II – Verifying Our Conclusion

- If enough women who did not have obvious signs of risk were left out to make the risk posed by each group proportionate, the number of women in our sample would decline by about 8.5%.

- In other words, women would make up 55.3% of our sample group, rather than 60.4%.

Part II – Verifying Our Conclusion

- This is a notable difference, and suggests that the company's current method for determining who is at risk either overpredicts the risk for women, or underpredicts for men.

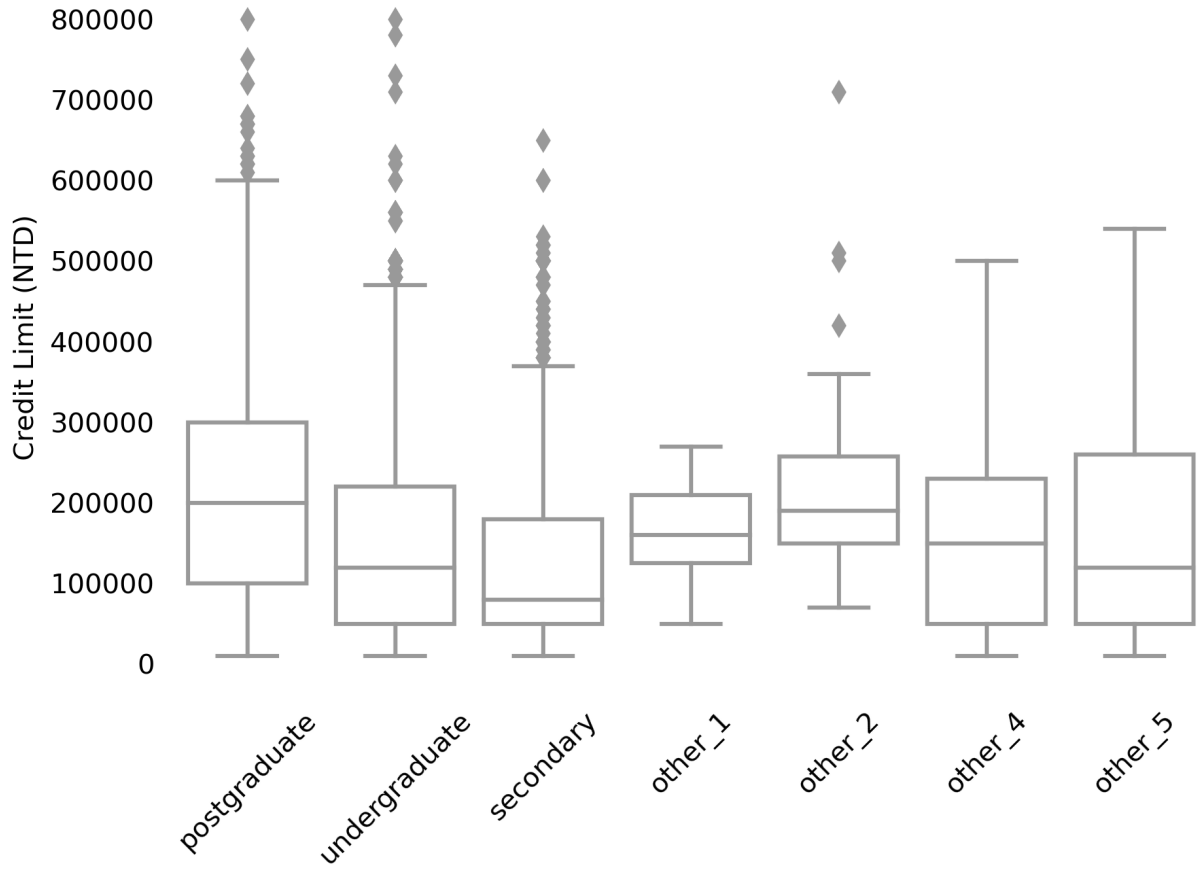
Part II – Additional Tests

- I ran several other tests to get practice with software, plotting, and running simulations.

- In these cases, tests were chosen not so much for inherent interest, but to see what happened when the data failed to meet some of the assumptions of the test.

Part II – Additional Tests

- For example, I tested whether there was a statistically significant relationship between education level and credit balance.
- This would normally require ANOVA.



Part II – Additional Tests

- But some of the assumptions of the f-test are not met.
- Distribution of values is not normal for several of the groups (large absolute value of *skew*).

EDUCATION	skew	var
postgraduate	0.623542	1.865420e+10
undergraduate	1.140935	1.461679e+10
secondary	1.492699	1.265015e+10
other_1	-0.207047	5.495238e+09
other_2	2.078121	1.396429e+10
other_4	0.877992	1.475478e+10
other_5	1.085159	2.896044e+10

Part II – Additional Tests

- There are also some notable differences in variance among the groups.

EDUCATION	skew	var
postgraduate	0.623542	1.865420e+10
undergraduate	1.140935	1.461679e+10
secondary	1.492699	1.265015e+10
other_1	-0.207047	5.495238e+09
other_2	2.078121	1.396429e+10
other_4	0.877992	1.475478e+10
other_5	1.085159	2.896044e+10

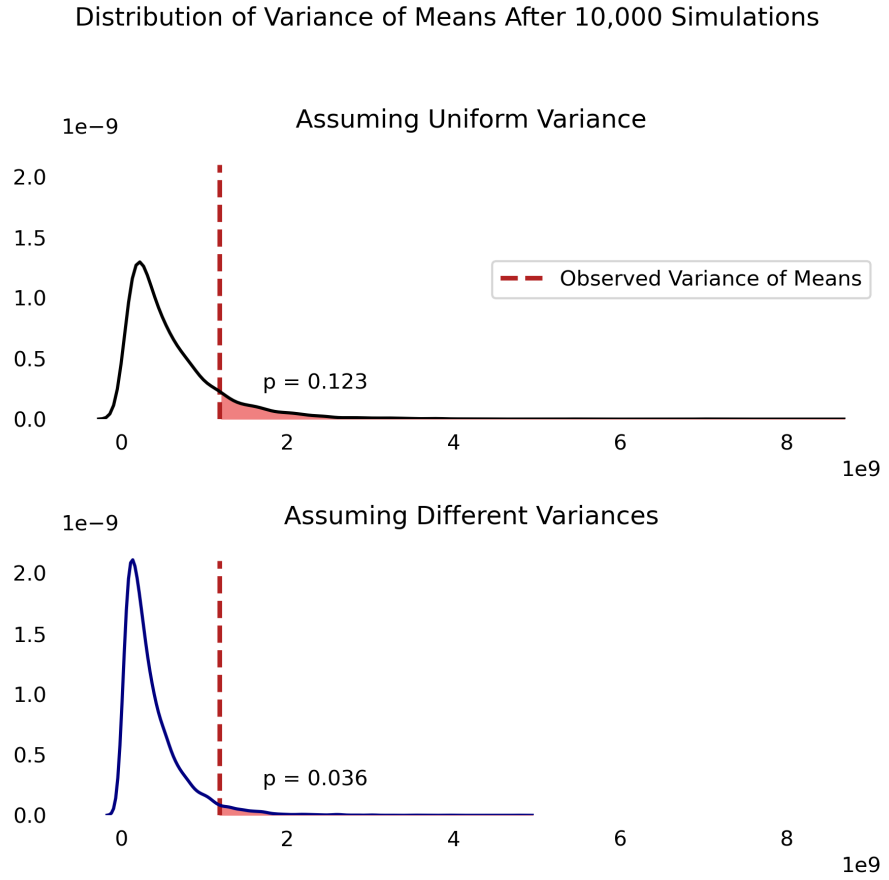
Part II – Additional Tests

- An f-test on the data produced an extraordinarily small p-value. (1.1e-161)
- But this was not reproduced in simulations.

EDUCATION	skew	var
postgraduate	0.623542	1.865420e+10
undergraduate	1.140935	1.461679e+10
secondary	1.492699	1.265015e+10
other_1	-0.207047	5.495238e+09
other_2	2.078121	1.396429e+10
other_4	0.877992	1.475478e+10
other_5	1.085159	2.896044e+10

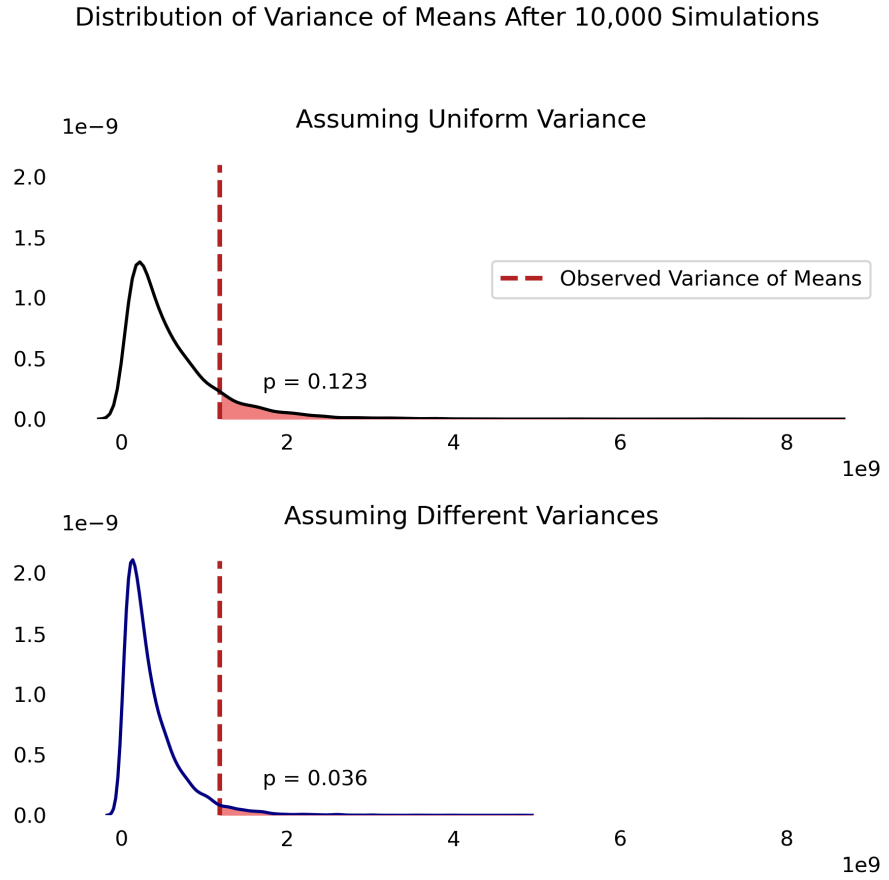
Part II – Additional Tests

- I ran two different simulations.
- One assumed variance among the different groups was uniform.
- The other did not.



Part II – Additional Tests

- In both cases p was vastly larger.
- Both also found the results to be not statistically significant.
- Remember we set α = .007.



Part II – Additional Tests

- The other tests and simulations can be found here: …

- This concludes the summary.