Executive Summary

Out of the 50,000 borrowers in this dataset, I estimate that 9,015 of them are going to charge off their loans at some point (including borrowers who already did). There is a slight prevalence of chargeoffs in the early days of a loan, though the chargeoff rate appears to stabilize around 150 days into the 3-day loan term.

I also postulate that 9,015 people may be a conservative over-estimate, as borrowers might be less likely to chargeoff if they are very near the end date of the three year loan. However, the data does not extend past 730 days into the loan period, so it is impossible to make any statistical claim about this hypothesis. This would be a topic of further exploration as more data is collected.

Technical Analysis

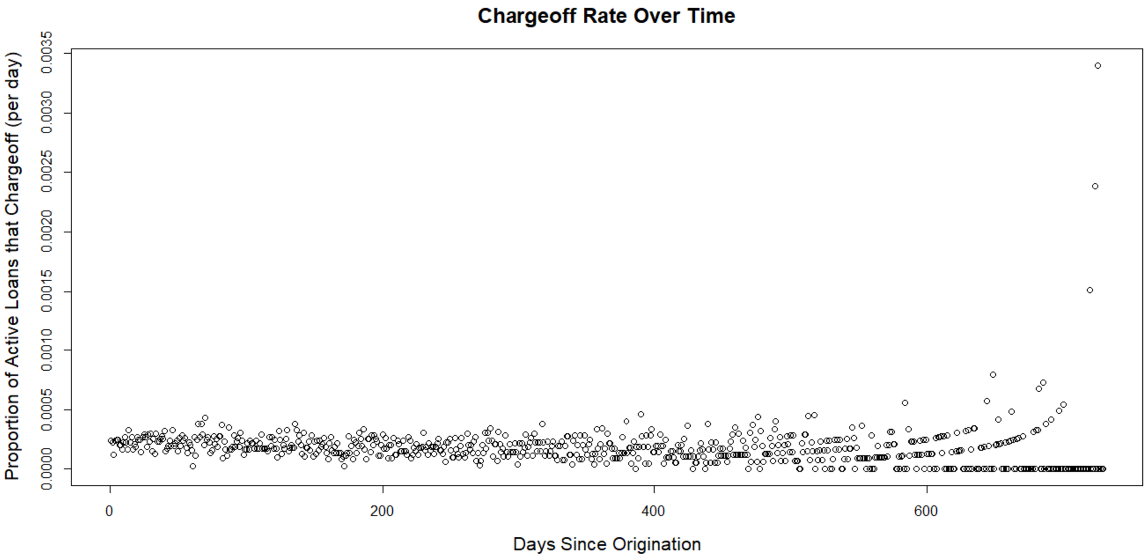
I first noticed that each borrower has a different amount of time remaining in their loan period, which yields a different probability of committing chargeoff for each borrower. For example, a borrower who's only 30 days into their loan has a greater risk of chargeoff than a borrower who's already 150 days into their loan. This intuition framed my whole analysis. For each borrower, I calculated a "chargeoff risk", measuring the probability that this borrower will charge off their loan:

$$\text{Chargeoff risk } = \begin{cases} 1 & \begin{array}{c}\text{if borrower has already charged off}\\ \text{at the time of data collection}\end{array} \\ p\begin{pmatrix}\text{chargeoff in days}\\ \text{X through 1095}\end{pmatrix} & \begin{array}{c}\text{if data is collected up to X days from}\\ \text{origination, with no chargeoff}\end{array} \end{cases}$$

Of course, calculating the "probability of chargeoff in days X through 1095" is the tricky part. To do this, I estimated the chargeoff rate for each day after origination, given by

$$\text{Chargeoff rate } = \frac{\text{Number of chargeoffs on day } X}{\text{Total number of active loans on day } X}$$

An "active loan" for day X refers to any borrower who has not yet committed chargeoff, and is still being observed X days into their loan period (no loan is observed for the full 1095 days yet in this dataset).



Note: The crazy variation at the end is not worrisome. As we increase the number of days since origination, we have less and less people to study: most people either committed chargeoff or simply haven't been studied for that long yet. As a result, we lose resolution on our measurements, as each borrower represents a larger proportion of the sample. Each dot within the rising tail indicates only a single borrower dropping out, but – because our sample size is dwindling by that point – that borrower represents a larger proportion of the sample, compared to individual borrowers early on. Again, this is a problem of resolution, not a true trend. Any worries are further alleviated when noticing that this period's average chargeoff rate perfectly matches other periods of average chargeoff rate (as seen in the upcoming table).

The dashed blue line signifies the first 20% of days counted from origination (days 1 through 146). This threshold is important because we see a higher proportion of chargeoffs per day during this early stage, compared to the rest of the dataset:

Days from Origination	0–20%	20–40%	40–60%	60–80%	80–100%
Avg. Daily Chargeoff Rate	0.00022	0.00018	0.00017	0.00015	0.00017

In this context, "100%" refers to the final day of current observation, 730 days after origination. We can see plainly that the 0–20% quantile has abnormally more chargeoffs than the other periods of observation. This suspicion is confirmed by running t-tests, binomial exact tests, and ANOVA tests (which are all covered in Appendix A). I proceed under the assumption that days 1–146 have an average daily chargeoff rate of 0.00022, while days 147–onward have a combined average daily chargeoff rate of 0.00017. This difference makes intuitive sense – the longer a borrower lasts without committing chargeoff, the chances become higher that they can reliably pay the rest of their loan as well.

Note: the difference between 0.00017 and 0.00022 might sound practically insignificant. However, a difference of 0.00005 would yield a difference in every 1 out of 20,000 people per day, or 2.5 extra people per day out of 50,000. That amounts to roughly 360 more people committing chargeoff in the first 146 days compared to any other 146-day quantile. When it comes to loaning money, that could be a lot! It's worth pointing out that extra attention should be given to new borrowers, versus established borrowers.

Using these calculations, we can compute that original "probability of chargeoff in days X through 1095", by summing over the daily probabilities of chargeoff for each individual day:

$$p\left(\begin{matrix}\text{chargeoff in days}\\ \text{X through 1095}\end{matrix}\right) = \sum_{X}^{1095} p(\text{chargeoff for each day})$$

where p(chargeoff for each day) is given by the average daily chargeoff rate for that specific day (0.00022 for days 1–146, or 0.00017 for days 147–1095).

For example, suppose we have data on one borrower up to 80 days after origination, and they haven't committed chargeoff yet. That individual's risk of future chargeoff would be

$$\begin{aligned} p\begin{pmatrix}\text{chargeoff in days}\\ \text{81 through 1095}\end{pmatrix} &= (146-80)\cdot 0.00022 \;+\; (1095-146)\cdot 0.00017 \\ &= 0.1759 \end{aligned}$$

Hence, I estimate that this borrower has a 17.59% chance of committing chargeoff, given that we have only tracked them for 80 days so far.

To estimate the total number of loans getting charged off, I simply calculate the expected value over all borrowers in the dataset:

$$\begin{aligned} \mathrm{E}\Big[\text{ Number of chargeoffs }\Big] &= \sum_{i=1}^{all\,people} p(\text{borrower } i \text{ commits chargeoff}) \\ &= 9015.297 \end{aligned}$$

Voila! We have arrived at my estimate of approximately 9015 chargeoffs.

Appendix

A Comparing Quantiles of Daily Chargeoff Rate

In this section, I will formally establish the difference between the 0–20% quantile (days 1–146) versus the other quantiles (days 147–730). Recall the following table:

Days from Origination	0–20%	20–40%	40–60%	60–80%	80–100%
Avg. Daily Chargeoff Rate	0.00022	0.00018	0.00017	0.00016	0.00016

Note that "100%" refers to the final day of observation, 730 days after origination. We can see plainly that the 0–20% quantile has abnormally more chargeoffs than the other periods of observation. I test this difference using Welch's two sample t-test at a 95% significance level:

Quantile Comparison	Mean Difference	p-value
Q1 vs. Q2	0.00004	$< .001$
Q1 vs. Q3	0.00005	$< .001$
Q1 vs. Q4	0.00006	$< .001$
Q1 vs. Q5	0.00006	0.028

The 0-20% mean is considered statistically significant from all other quantiles. If I controlled for multiple testing error, the last quartile might fail to show significance due to its larger variance (as discussed in the Technical Analysis), but - regardless - it seems pretty clear in context that the 80–100% quantile should probably fall in line with the 60–80% quantile.

Moreover, I further confirm my suspicions by running a binomial exact test: if the daily chargeoff rate was actually the same within each quantile, then a daily chargeoff rate from the 0–20% quantile should have a simple 50% chance of exceeding a daily chargeoff rate from any other quantile. This yields a binomial distribution, with $n = 146$ and $p = 0.5$. I compare day X of the 0–20% quantile against day X of the other quantiles, comparing them over all values of X from 1 to 145, yielding the following results:

Quantile Comparison	# days where Quantile 1 > Quantile N	p-value
Q1 vs. Q2	94	$< .001$
Q1 vs. Q3	101	$< .001$
Q1 vs. Q4	97	$< .001$
Q1 vs. Q5	112	$< .001$

Quantile 1 clearly demonstrates a preponderance of higher chargeoff rates, so we can safely conclude that early days of a loan period contain higher risk of chargeoff than later days.

To see whether any further stratification of quantiles is needed (beyond separating the first quantile from the rest), we conduct a one-way ANOVA on the remaining four quantiles:

Hypothesis:	no significant difference in the average daily chargeoff rate between quantiles 2–5
p-value:	$p = 0.606$ (insignificant)
Conclusion:	Reasonable to equate the average daily chargeoff rate between quantiles 2–5

I thereby justify the approach of treating the 0–20% quantile separately, while pooling together the remaining 20–100% quantiles.

B Future Directions

I extrapolated the average daily chargeoff rate from days 146–730 outward to days 731–1095. However, it would be reasonable to hypothesize that people might not commit chargeoff if they are very close to finishing their payments, due to successful history and a “being so close” or “made it this far” attitude. This would be a topic of further exploration as more data is collected in the future.

There are also almost certainly factors that might predict individual chargeoff probabilities: age, income, location, and other demographic qualities may be associated with chargeoff rates. It would be worthwhile to follow up on this project by gathering this data and exploring any associations.