Clipboard Health Price Case Study

Christopher Ian Williamson Guerrero

Relevant Information:

Gmail: Christopher.i.williamson.guerrero@gmail.com

LinkedIn: https://www.linkedin.com/in/christopher-williamson-guerrero/?locale=en_US

GitHub: https://github.com/CWG-DS/Profit-Optimazation

Introduction

The end goal of this project is to determine the most optimum pricing strategy for the case presented to us by Clipboard Health. We begin by dissecting the case study presented to us.

1. The launch of our ride-hailing service will be active for 12 months.
2. Riders are charged $30 per ride.
3. Driver's pay per ride is our main variable to be determined. Drivers are able to choose which rides to service based on compensation. We are given an extract that describes Driver's decision making based on payment.
4. There is a total of 10.000 Riders which are available to us.
5. We are limited to an increase of 1.000 drivers per month.
6. The number of rides that are requested by Riders is determined by a Poisson distribution with and initial Lambda of 1.
7. Riders who do not request rides or do so but are not serviced by Drivers exit the program and do not return.
8. The number of requests per driver is determined by a Poisson distribution with an initial Lambda of 1. Subsequent requests performed by Riders which stay in the program will also exhibit a Poisson distribution with a Lambda equivalent to the number of serviced rides from the previous month.

Based on these restrictions we will proceed to model our data and output the most efficient payment method to Drivers so as to maximize profits.

Data Exploration

The sample of data we are given is composed of 1.000 data points. It describes Driver's acceptance of ride requests based on the amount of pay they will receive. An example of the data is presented below:

	PAY	Accepted
0	29.358732	0
1	22.986847	0
2	18.020348	0
3	45.730717	1
4	14.642845	0

The [PAY] column refers to the amount offered to a Driver whilst the [Accepted] columns codifies whether the offer was Accepted [1] or Declined [0].

In order to better visualize our data, we divided our sample based on whether the PAY amount was Accepted or Declined through the use of boxplots as shown in Figure 1.

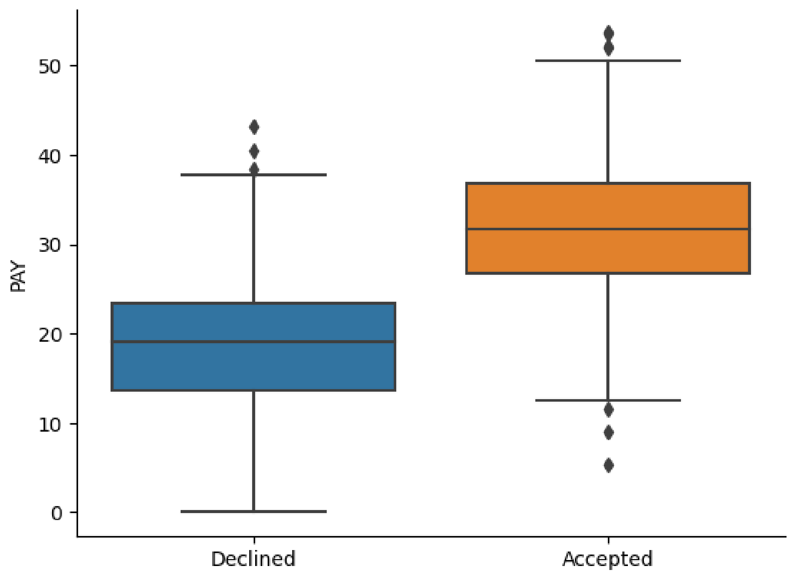

Figure 1: Box plot describing the distributions of Declined and Accepted requests per driver payment range.

As it would be expected, the distribution pertaining to Declined requests is characterized by a lower pay average ($\mu = 18.62$) when compared to Accepted requests ($\mu = 32.08$). Additionally, the boxplots show us the existence of outliers which might skew further analysis. Thus, it was decided to discard possible extreme values through the use of the interquartile method. Any value which exceeded 1.5 times the interquartile range was discarded. After which, we tested for normality through the use of the Shapiro-Wilk Test determining that both distributions were in fact Gaussian/Normal distributions.

We then proceeded to collapse both distributions into a single plot which describes the probability of rejection for every driver payment point. As we can see in Figure 2 there is a reduction in the probability of declined rides as we increase the pay offered to drivers. However, even with a low sensitivity (as described by using large intervals, $1 in our case), we do not have enough data to plot a smooth transition between intervals. As we can see, there is noticeable sudden spikes within the (31, 32] and (37, 38] intervals, as well as no values within the (2,3] interval. In order to solve these problems, we increased our sample size by generating normal distributions with the same characteristics as our initial ones.

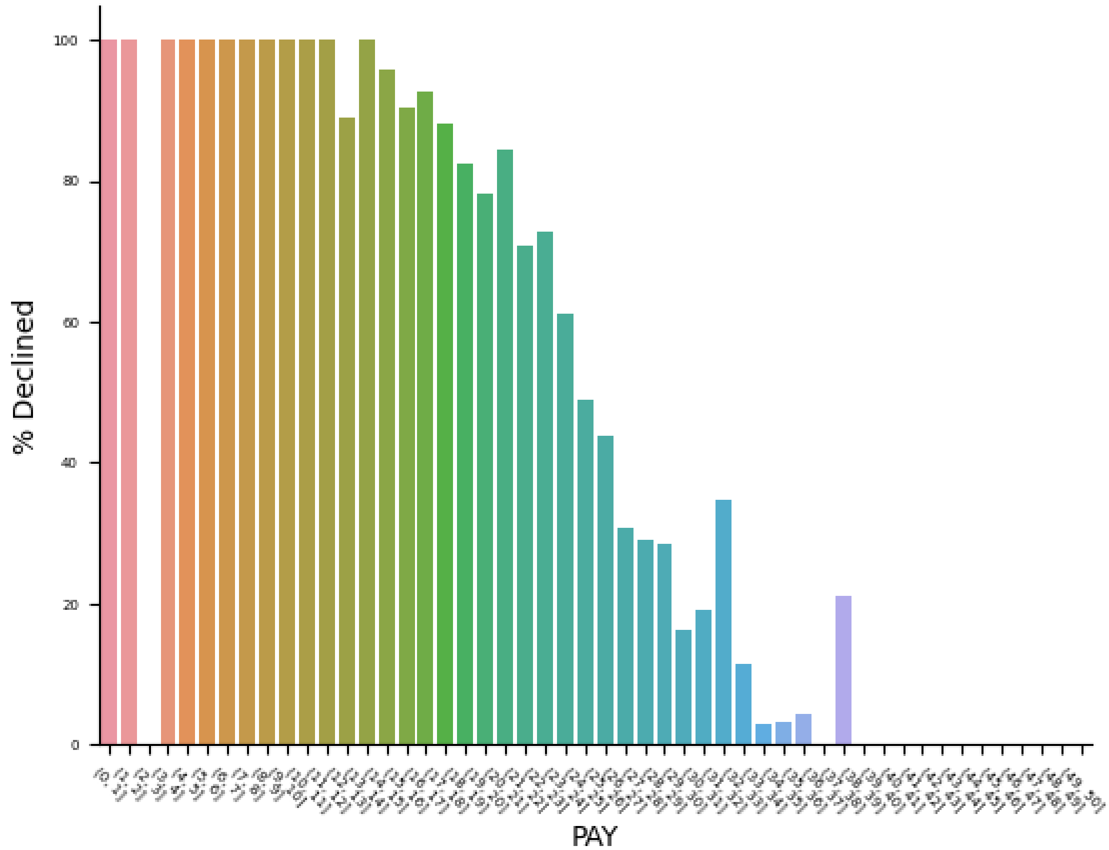

Figure 2: Bar plot showing the probability of a request being declined per driver payment interval. Intervals are 1$ wide.

Data Generation

Given that both Decline and Accepted distributions both exhibit a Gaussian/Normal distribution we were able to generate further data points based on the characteristics (mean and standard deviation) of their respective distributions. We will therefore recreate these distributions with a sample size of 100 million data points per condition. This increase in our sample size will enable us to generate a more accurate estimate of which rides were accepted/decline for every driver payment range as well as increase our sensitivity by reducing the width of our intervals to $0,01 instead of the previous $1 range. It is of note that values bellow 0 were discarded. The total number of values discarded per condition is 0.05% for Declined and < 0.001% for Accepted. No effects are expected from the removal of such a small portion of the sample. Both newly generated distributions are shown in Figure 3.

We then proceeded to collapse our newly generated distributions into a single plot much like we did before. As seen in Figure 4, our resulting percentage of declined rides describes an inverse sigmoid distribution. We can also appreciate some deviance from this distribution at the right tail end caused by extreme values. However, since $30 marks our break-even point it is irrelevant for our specific case and will have no effect on further analysis.

Now that we have our data cleaned and sorted, we will proceed with the pricing strategy.

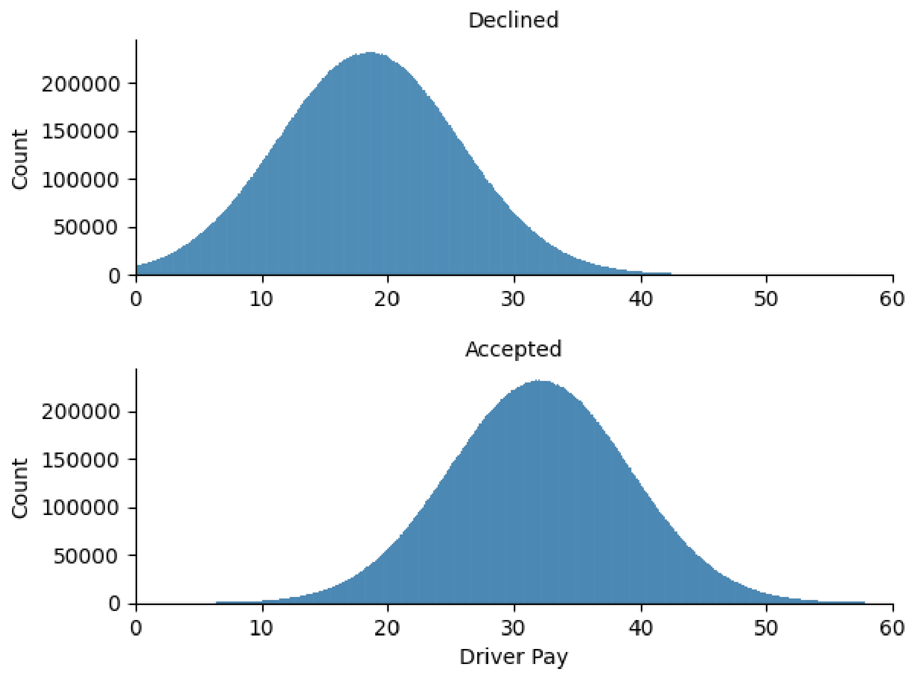

Figure 3: Newly generated distributions per Declined/Accepted conditions with Driver Pay in the x-axis and number of occurrences in the y-axis.

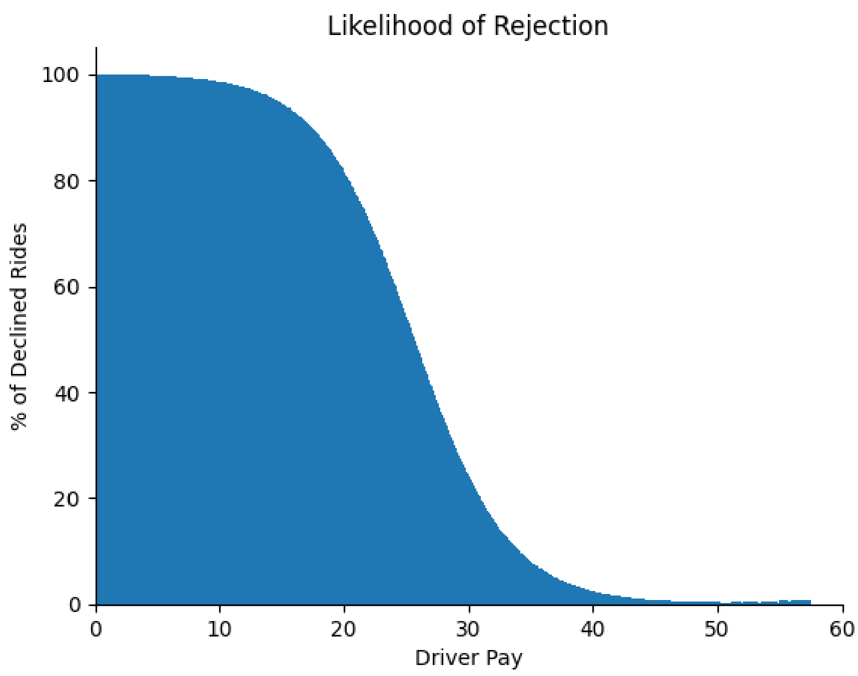

Figure 4: Inverse sigmoid distributions which describes declined requests probability for every driver payment point.

Fixed Pricing Strategy

We first focused on a fixed pricing strategy where driver pay is fixed throughout the whole duration of our program. To do so we employed a custom function which outputs the total profit acquired during the 12-month span of our program given a driver payment amount as input.

The function logic can be found in the Readme.md document within the following link: https://github.com/CWG-DS/Profit-Optimazation and the function it self with detailed comments describing what each line of code does can be found in the Clipboard_Health.ipynb within the same repository.

We proceeded to input every driver pay point from $0,01 to $30,01 with $0,01 increments to our function and plotted the results as shown in Figure 5.

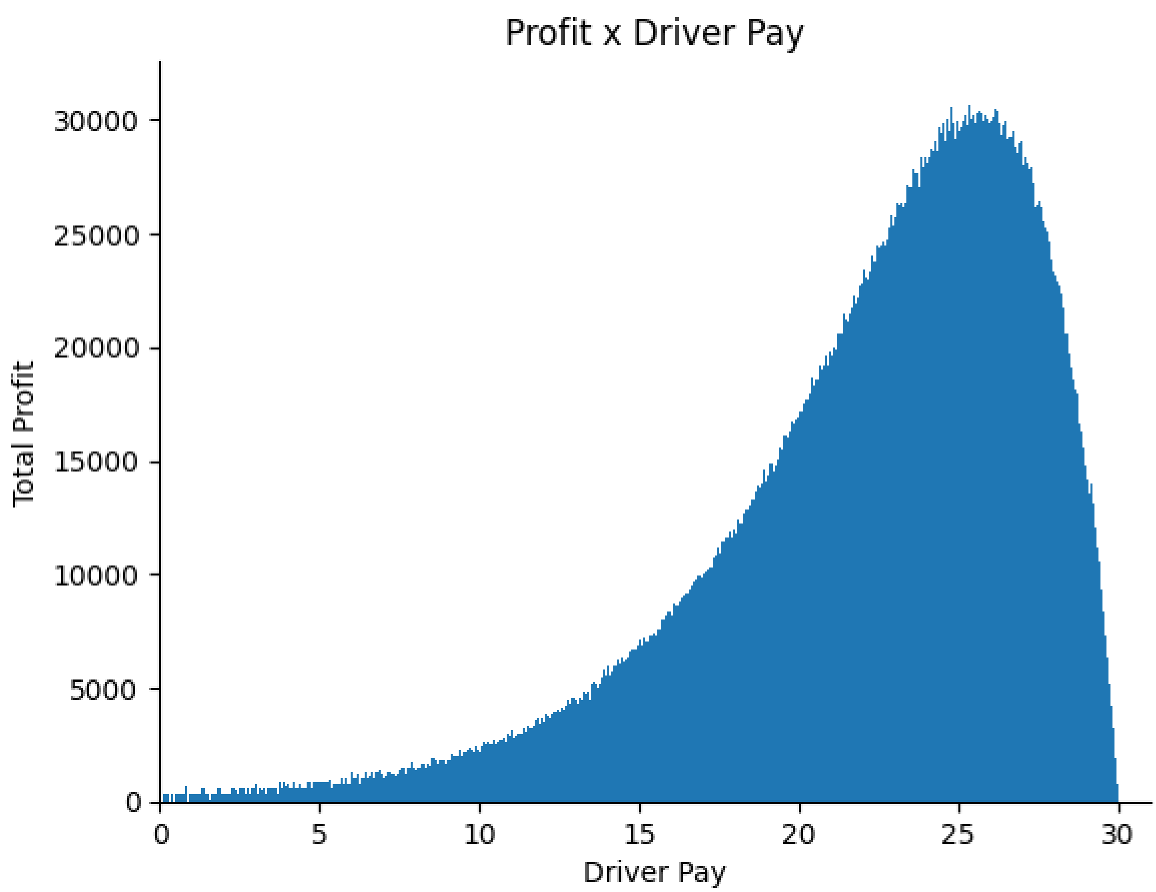

Figure 5: Total profit achieved per driver payment point.

As shown in Figure 5 there is a progressive increase of total profits up to a certain payment point. However, this data is extracted from a single iteration per price point and therefore, while the overall trend might be accurate, individual profit values per price points might fluctuate. In order to solve this, we can further iterate over a smaller range of possible driver payment amounts so as to achieve greater accuracy. To do so, we set a $29.000 cut off point and iterated our function 100 times over the resulting interval [24.15, 26.82] payment range.

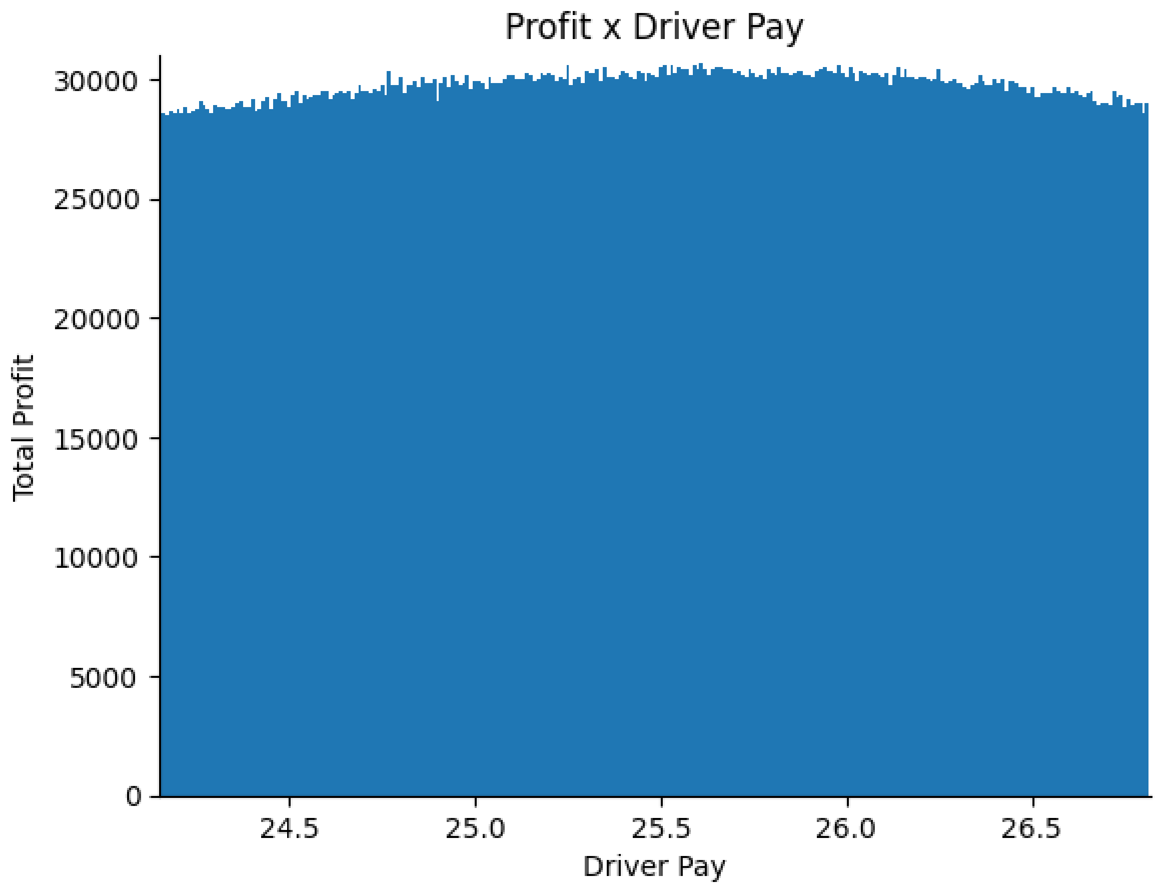

Figure 6: Total profit generated per driver payment point within the interval [24.15, 26.82] $.

As we can appreciate in Figure 6, total profits become quite stable within the stated range. The average expected profits within this range is $29.758,48. When tested for normality through the Shapiro-Wilk Test the distribution was determined to be non-normal. Thus, one of two approaches should be taken to determine confidence intervals, resampling the data so as to obtain a normal distribution or via bootstrapping methodology. However, this was not done in this project and will be revisited in the future.

It is then concluded that the most optimum driver payment range using a fixed priced strategy is anywhere between $24,15 and $26,82 with an expected total profit of approximately $29.758. Greater profits might be achieved by tightening the driver pay range, however the variability already present within this range makes further reduction questionable.

Variable Pricing Strategy

We then set to explore the possibility of applying a variable pricing strategy where initial losses would be accepted in exchange for a greater growth of the user base, after which, driver pay would be reduced so as to offset these initial losses and take advantage of an increase of rides requested.

To do so we set to explore the behavior of our userbase under four driver payment amounts. A value within our maximum profit interval using the previously discussed fixed pricing approach, $25. Our break-even point, $30. A value equidistant over the breakeven point, $35. And an extreme value, $40.

These amounts would give us a good understanding of how the userbase interacts with our service over different price points and how will a change in pricing affect said interaction. To do so we focused on three different, but related, variables: Number of accepted rides per month, percentual increase in rides per month and, profit generated per month. For a variable pricing strategy to achieve a greater performance than our previously discussed strategy it would have to significantly outperform our fixed pricing strategy by increasing the number of accepted rides and maintain them to an extent when the pricing change occurred so as to offset the initial losses incurred.

To examine this possibility, we set a baseline by modifying our previous custom function to output our desired variables.

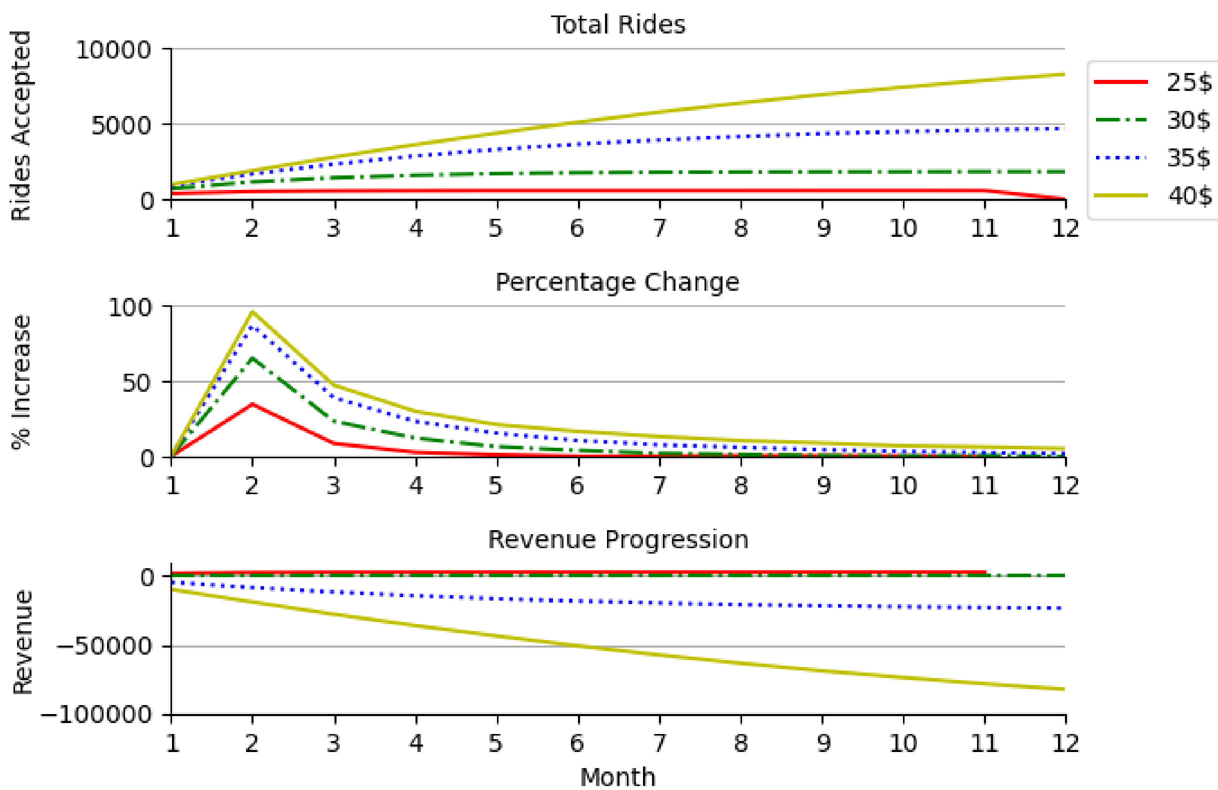

Figure 7: The image depicts three separate subplots. Our top subplot depicts the number of accepted rides per month depending on how much drivers were offered. The middle subplot shows the degree of change in the number of rides accepted with respect to the previous month as a percentage. The bottom subplot indicates the profit, or loss, achieved by month.

As expected, there is an increase of rides accepted as we increase the amount of payment which drivers receive. However, the progressive increase in the number of rides accepted does eventually taper off. The rate of the decrease in growth per month also depends on the amount paid to drivers. Both these effects are strongly linked to our inverted sigmoid distribution which dictates the percentage of declined rides. It is of note that the biggest growth in our userbase occurs within the initial months. This factor, in addition to the great increase in losses incurred, as shown by our final subplot of driver payment points above our breakeven point, suggests that a price change should occur between the second and forth month so as to minimize the impact of losses while maximizing user growth.

We then replicated our analysis with a slight modification. When the function arrived at the fourth iteration the driver payment amount changed to $25, an amount which has previously shown to maximize profits.

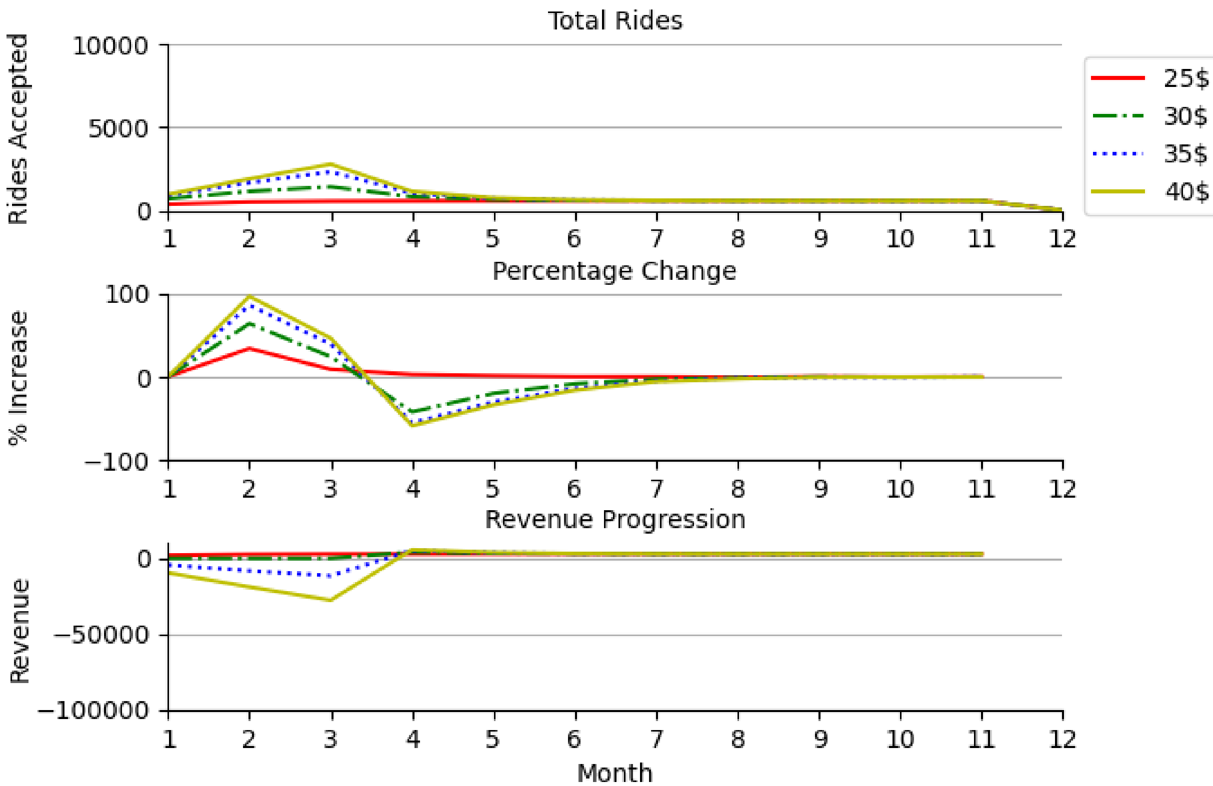

Figure 8: The figure shows the same three previous subplots, as shown in figure 6, with the exception of a price change occurring in the fourth month.

Our results are shown in Figure 7. The resulting figure is identical to the former during the initial 3 months. However, we can clearly see the effects of the pricing change with a rapid decline of rides accepted which rapidly converges with the $25 plot. Our third subplot shows an alarming insight, with price reversal not showing a proportional gain in profit with regards to the losses incurred during the initial months. This trend points towards the unfeasibility of employing a variable pricing strategy as the user attrition rate is not proportional to the gain during the initial months. After further analysis varying the month of implementation of our payment change, we are able to conclude that the earlier the change is established the better the outcome. However, the pattern shown in figure 8 remains with no combination of implementation and payment over the breakeven point being able to outperform our fixed pricing strategy.

Conclusion

After a detailed exploration of the data and conditions given to us by the problem, we conclude that the most profitable strategy is a fixed pricing strategy where drivers are paid an amount between $24,15 and $26,82 with an expected result of $29.758 in profits. Additional analysis using bootstrap methodology will be able to give us a confidence interval to our results.

Replication of this project might find deviations in the most optimum driver payment interval within the cent range due to differences within the data generation stage. For future projects a seed will be attached to the data generation function in order to enable more accurate replication of our findings.