Achieving Economic Goals:

Driver Lifetime Value & Churn Factors

By Alyssa Thy Bui

Content

Problem Framing

Problem Framing

Context

A ride hailing service’s success heavily depends on effectively matching supply and demand, with a robust and engaged driver network playing an essential role in achieving this balance. Specifically, understanding the factors that contribute to a driver's loyalty and value to the platform is crucial for sustainable growth and operational efficiency.

Business Goal

Optimize Driver Retention and Value for Revenue Growth

Report Objectives

1. Uncover trends and insights regarding driver’s engagement on the platform.
2. Recommend driver Lifetime Value (LTV).
3. Discover determining factors that influence driver churn.

Key questions to answer:

1. How to determine a driver’s lifetime value to the platform?
2. What were the determining factors of driver retention?
3. Which driver segments were at risk of churn and what can we do about it?

Methodology

Uncovering trends

Methodology

1. **Data Preparation** *(Appendix)*
   - Combine driver onboarding data with ride data into a master table, ensuring proper joins.
   - Clean and preprocess the data to handle any inconsistencies.
2. **Summary Statistics**
   - Calculate key metrics such as:
     - Total number of active drivers and rides
     - Total ride values
3. **Engagement Metrics by Week**
   - Create a weekly breakdown to visualize these metrics to show trends over time:
     - Number of active users (drivers who completed at least one ride)
     - Total number of rides completed
4. **Cohort Analysis for Retention**
   - Group drivers into cohorts based on their onboarding week.
   - Calculate retention rates for each cohort at every one week after onboarding.
   - Create a cohort heatmap to visualize retention patterns.

LTV Calculation

Methodology

1. **Define the timeframe** for the calculation (ex, 1 year, 2 years, average lifespan, etc.)
2. **Identify key components**
   - Revenue: Total earnings generated by the driver
   - Costs: Expenses associated with the driver (e.g., onboarding, support, incentives)
   - Churn Rate: How likely driver stays with the platform after the timeframe defined.
3. **Recommend a formula** that incorporates key components given the provided data
4. **Analyze factors affecting LTV**
   - Conduct multiple regression analysis to study the relationship between LTV and other potential factors such as:
     - Average ride distance
     - Average ride duration
     - Percentage of prime time rides
     - Frequency of ride

Churn Analysis

Methodology

1. **Define churn**
2. **Investigate factors affecting churn** through the following approaches:
   - Hypothesis Testing: Formulate hypotheses and conduct statistical tests (e.g., chi-square, t-tests) to validate.
   - Others (if time allowed):
     - Logistic regression model: Analyze coefficients to understand the impact of different features on churn.
     - Tree-based classification model: Evaluate feature importance to identify key drivers of churn.
3. **Segment drivers to identify high-value drivers at risk of churning** using RFM
   - Define and calculate RFM metrics.
   - Divide each metric into three levels.
   - Combine RFM score for each driver.
   - Define different segments.
   - Identify high risk drivers and recommend strategies.

Insights

Uncovering trends

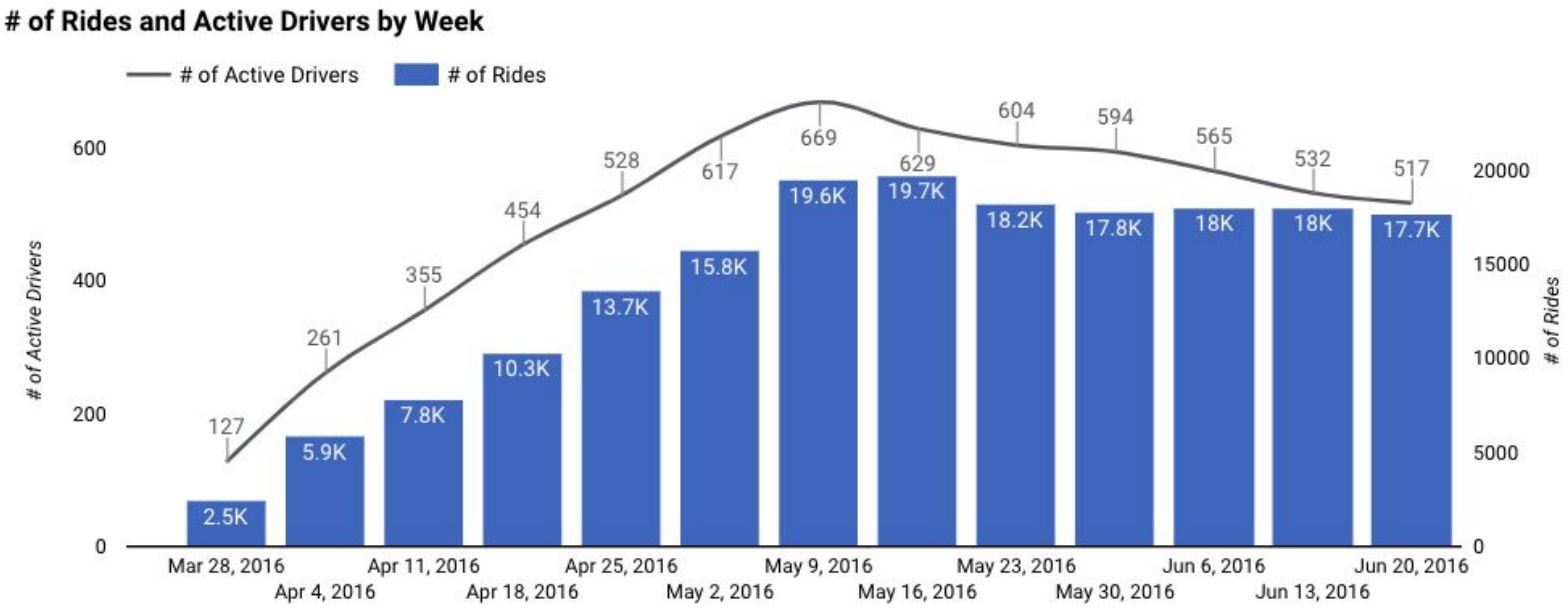
> Breakdown by Week

90-day	193,502	937	$ 2,619,345
Timeframe	# rides	# active drivers (who completed at least one ride)	ride value

1. The number of rides positively corresponded with the number of active drivers.

2. After the week of May 9, despite no data on new onboarding drivers, the number of rides remained stable afterwards, indicating that business performance was driven by loyal drivers, which further emphasized the importance of retaining drivers and maximizing their values.

3. Based on the overall stats, **the average ride value was $13.54** and **the average number of rides per day per driver was 2.3.**



Uncovering trends

> Cohort Analysis

1. After onboarding, there seems to be a natural drop-off in driver engagement, with approximately 83 drivers not completing any rides after signing up.

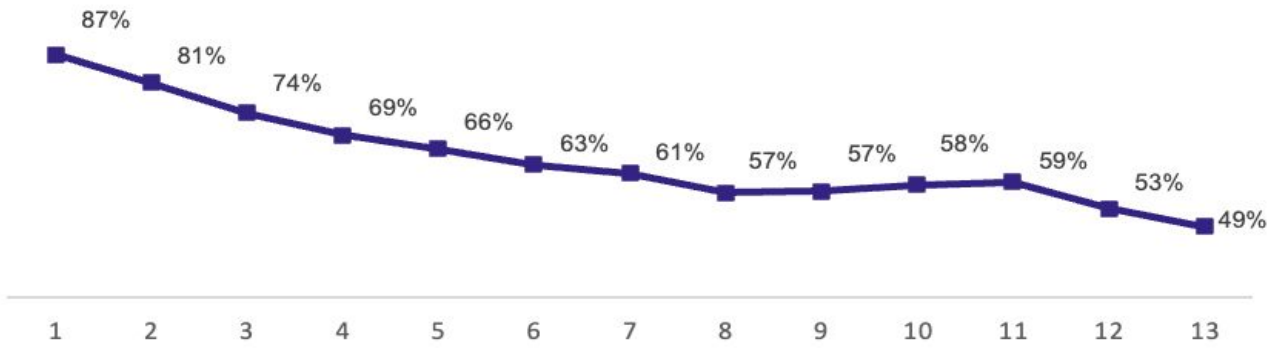
2. The retention rate decreased gradually over the week, with relatively similar trends over each onboarding cohort.

3. On average, the driver retention rate by week 12/13 was around 50%, indicating that the median churn point for drivers typically occurred between weeks 12 and 13. **This is the average lifespan of a typical driver.**

Retention rates by week after onboarding

Onboarding Week	New Drivers	1	2	3	4	5	6	7	8	9	10	11	12	13
2016-03-28	138	91%	82%	70%	68%	64%	62%	60%	57%	54%	57%	56%	51%	49%
2016-04-04	163	90%	80%	77%	69%	71%	63%	62%	61%	63%	63%	61%	55%	
2016-04-11	142	88%	81%	71%	71%	65%	66%	63%	60%	59%	58%	60%		
2016-04-18	137	86%	84%	79%	74%	66%	62%	65%	55%	57%	55%			
2016-04-25	125	86%	79%	73%	67%	68%	65%	59%	51%	51%				
2016-05-02	121	86%	83%	74%	69%	65%	63%	57%	55%					
2016-05-09	111	85%	79%	77%	68%	66%	59%	60%						

on average



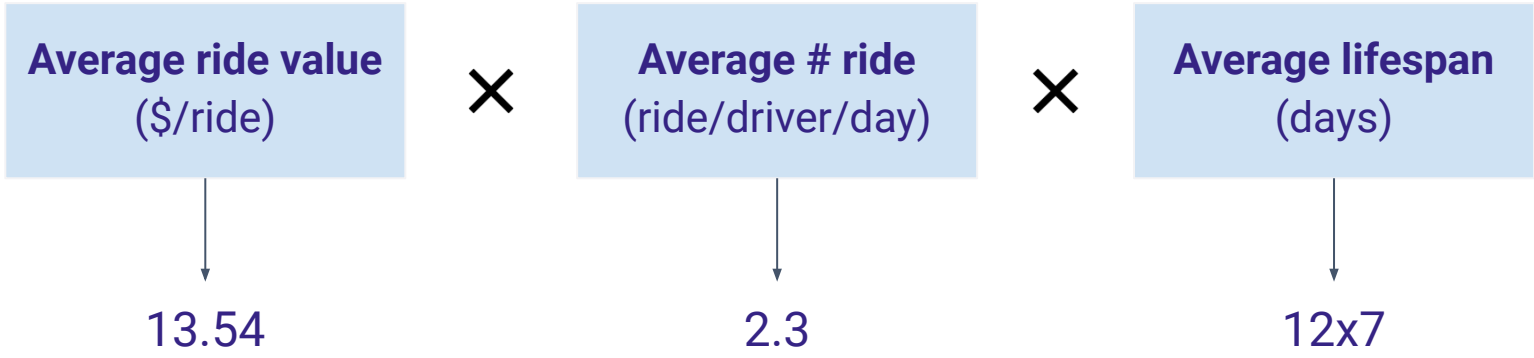
LTV Calculation

> Formula

Driver Lifetime Value (LTV) informs businesses of the total revenue expected from a driver over their entire tenure, guiding decisions on onboarding, retention strategies, and resource allocation.

Understanding LTV helps businesses optimize profitability by focusing on retaining drivers and maximizing their long-term revenue potential.

With the previous information, **Driver Lifetime Value** can be calculated using the following formula:



Driver LTV = $ 2,615.93

Note: For simplification purpose, costs to acquire and retain a driver were not included in the calculation because these costs varied in different regions and across different timeframes.

See Appendix for a full list of assumptions of LTV.

See Appendix for the assumptions of ride fare calculation.

LTV Calculation

> Determining Factors

To better understand the factors affecting driver LTV, we will examine LTV at an individual level, identify potential factors, and conduct a multiple regression analysis.

A multiple regression is a statistical method that examines how several different factors together can influence a specific outcome.

Equation

LTV = **−1408.57** + **0.109** × avg distance per drive + **1.375** × avg duration per drive − **2.511** × avg arrival duration + **80.019** × avg wait duration + **392.022** × ride freq during active + **25.002** × avg primetime multiplier

Key Insights

After running a multiple regression model for six potential factors, here are the implications:

- Ride frequency during active periods is the strongest predictor of LTV.
- Longer wait durations are positively associated with LTV, which is counterintuitive and may need further investigation.
- Prime time rides contribute positively to LTV.
- Longer arrival durations negatively impact LTV.

See Appendix for a detailed model summary

While further investigation is necessary before reaching definitive conclusions, it is evident that implementing engagement strategies to increase ride frequency and developing retention plans to extend drivers' tenure will enhance driver value, thereby driving revenue growth for the business.

Churn Analysis

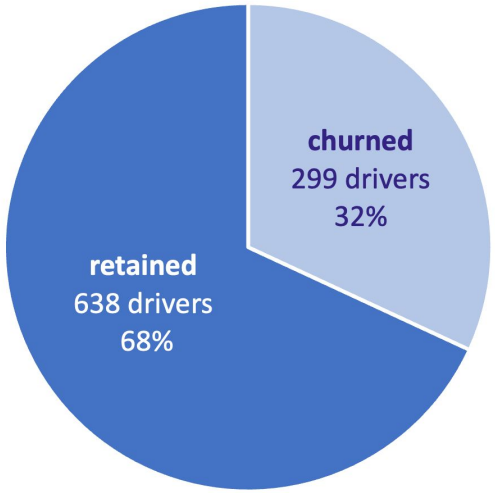
> Defining Churn

In an effort for retention, a churn analysis is needed to establish a clear definition of churn and identify patterns or predictors that can help mitigate it effectively.

Definition of Churn

After trying out different approaches, the most appropriate definition of churn based on the provided data and business sense is as follows: **A driver is categorized as "churned" if they have not completed any rides within the past three weeks or longer.**

and % of retained vs churned drivers among 937 drivers having ride data



See Appendix for details on the chosen approach.

Churn Analysis

> Determining Factors

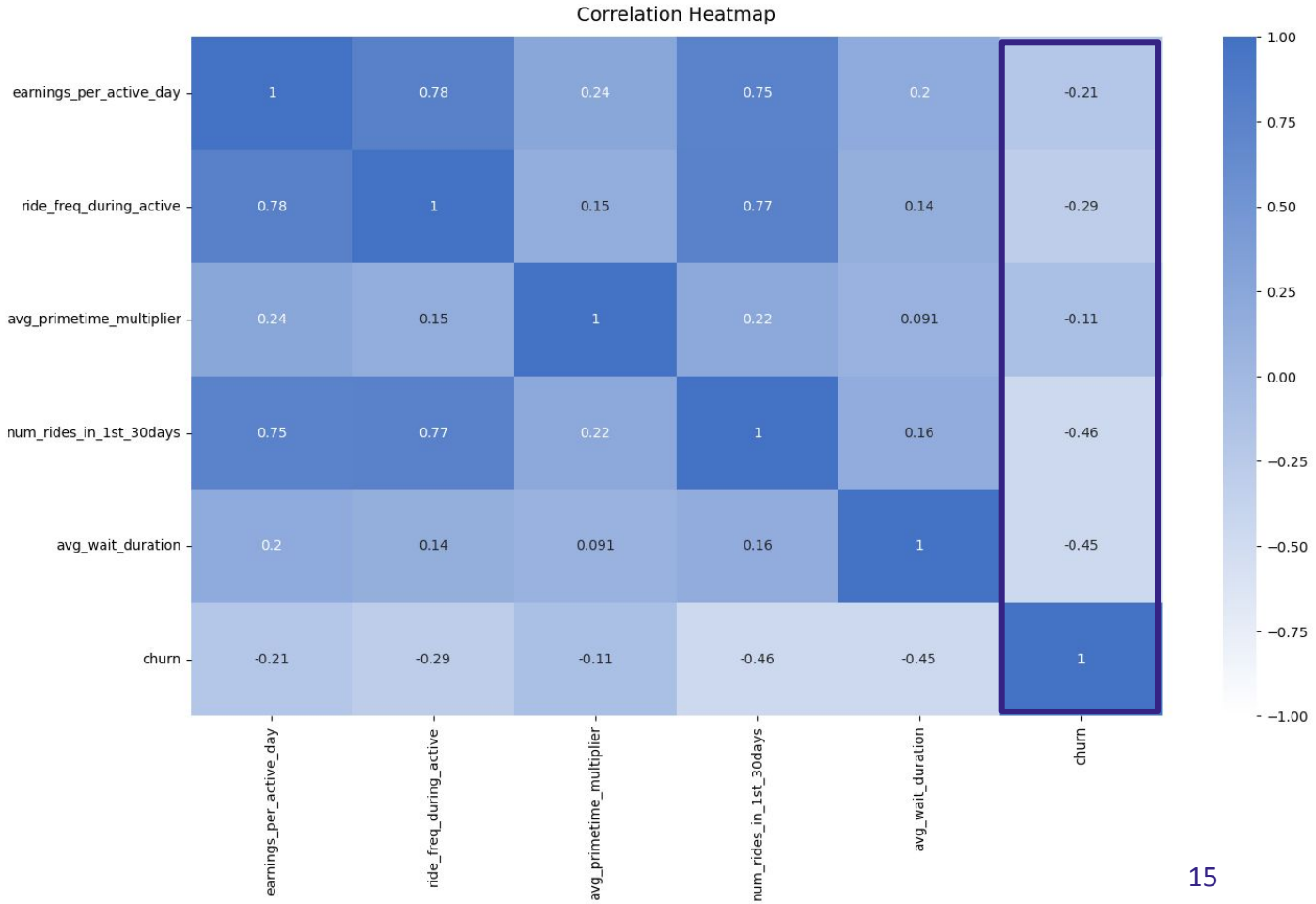
Hypothesis to be tested for the potential determining factors:

1. Earnings per active day
2. Ride frequency during active period
3. Exposure to prime ride
4. Number of trips taken within the first 30 days.
5. Average wait time when picking up customers.

See Appendix on how to calculate these factors.

Correlation

First, we will check the correlation to identify which variables are most strongly associated with churn. Among the five, ***avg_wait_duration*** and ***num_rides_in_1st_30days*** had moderate negative correlations.



Churn Analysis

> Determining Factors

Hypothesis to be tested for the potential determining factors:

1. Earnings per active day
2. Ride frequency during active period
3. Exposure to prime ride
4. Number of trips taken within the first 30 days.
5. Average wait time when picking up customers.

See Appendix on how to calculate these factors.

Statistical Tests

Then, we will validate these predictors through conducting statistical tests.

- The null hypothesis states that the mean value of each feature was the same for churned and retained drives. The alternative hypothesis states that the mean was different for the two groups.
- The significance level is set at 5%.

The statistical t-tests indicate that there was a significant difference in the means of each feature, thereby confirming the hypothesis. *See Appendix for detailed results.*

Conclusion

Through different approaches, **ride frequency** was one of the strongest predictors of churn. Knowing this information will help business design appropriate strategies. More **exposure to prime rides** as well as incentives or bonuses will **increase earnings** for each driver, motivating them to be more active. Besides, it's important to engage new drivers during **their first X days mark** since numbers showed early active users were more likely to retain. More analysis is needed to determine the exact value for this golden window to fully convert new drivers.

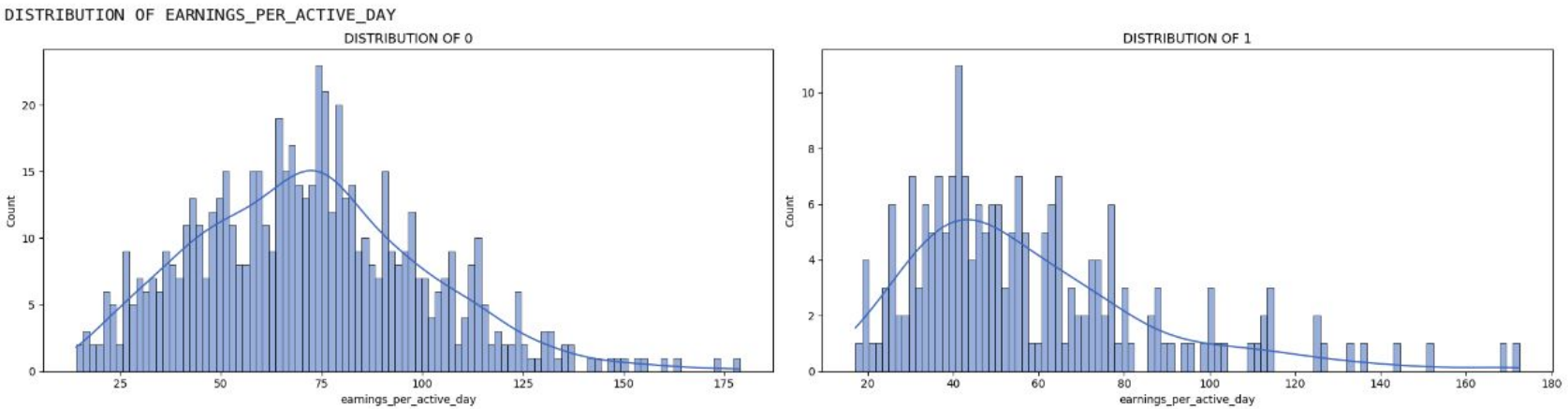
We also need to investigate the counterintuitive relationship between longer average wait time and lower churn/higher LTV.

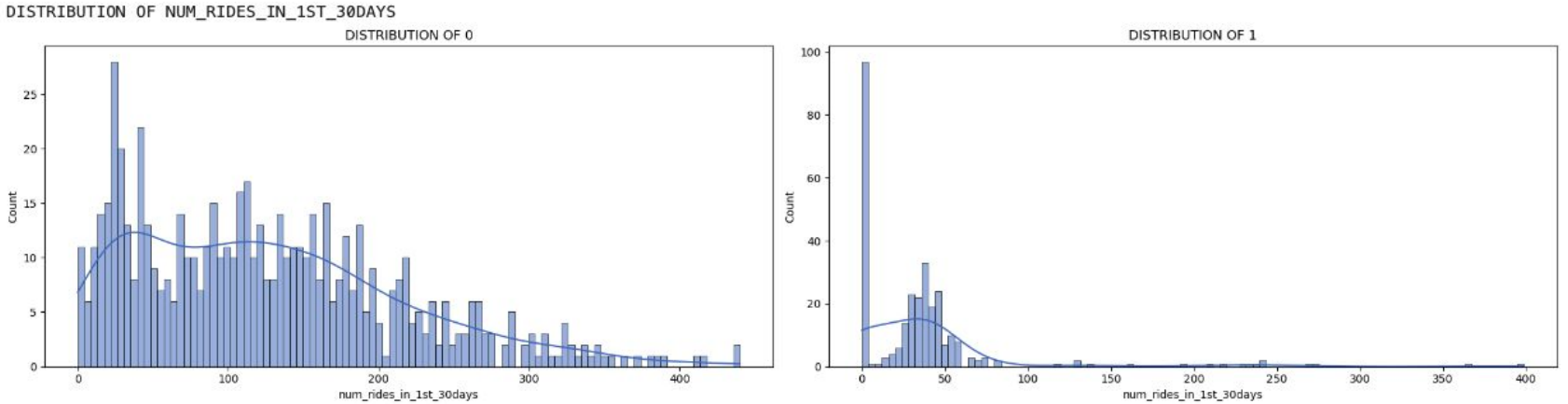
Churn Analysis

> Determining Factors

Visualization

The following charts show the differences in feature distribution between the Churned (1) and Retained (0) groups. It can be clearly observed that the engagement level between two groups were quite different for all features. *See Appendix for all 5 features' distribution.*





Churn Analysis

> Segmentation

RFM (Recency, Frequency, Monetary) is a robust segmentation method because it provides a comprehensive view to identify and target high-value drivers more effectively based on their past engagements and earning patterns.

From these segments, we can identify drivers who can generate significant value for the company but exhibit early signs of churn, in order to prioritize targeted retention efforts and maximize their long-term retention.

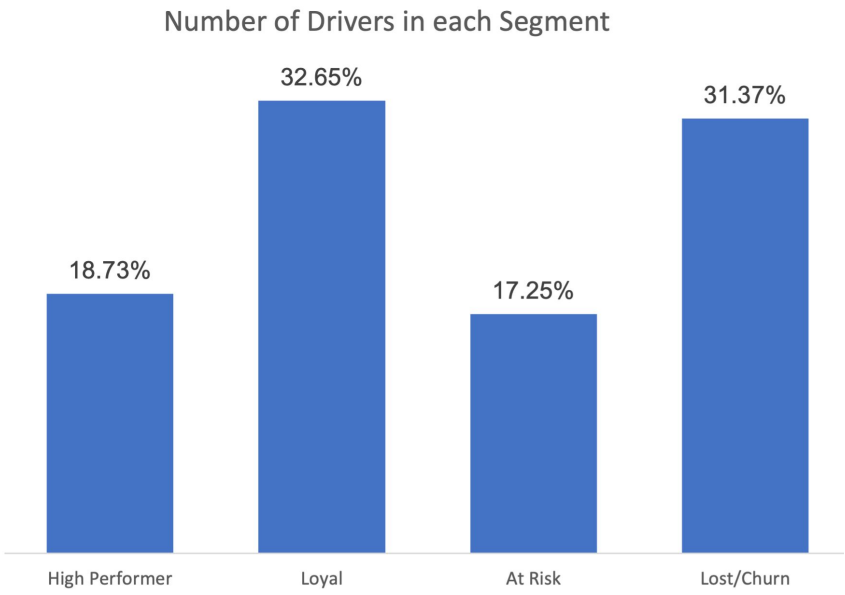
Driver Segmentation based on RFM

High performer: High value, frequent drivers

Loyal: Consistent drivers

At risk: Previously engaged but showing decreased activity (*See Appendix for more details on this segment*)

Lost/ Churn: Had not driven in more than 1 month



See Appendix on details of segmentation process.

RFM Metrics *Appendix*

Recency: Number of Days Since Last Drive

Frequency: Ride Frequency During Active Period

Monetary: Ride Values per Active Day

Based on the given data, segmentation is limited to the RFM (Recency, Frequency, Monetary) model. With additional data such as demographics and app interactions, a more comprehensive approach could identify likely churn customers, considering factors like seasonal or low-tech drivers.

Recommendation

Conclusion

Recommendation

In conclusion, it's highly recommended to closely monitor the retention trends of drivers to enable timely actions, as this directly impacts a driver's lifetime value and, consequently, the cash flow. Particularly, the "At Risk" segment poses a significant challenge for the company; these drivers, who have the potential to generate substantial value but show early signs of churn, display the most unpredictable behavior.

In addition, with ride frequency and earnings being the strongest predictors of churn, there are a few initiatives to consider:

- **Personalized incentives or bonuses:** Design a reward program for completing a certain number of rides or for receiving high ratings.
- **Increased exposure to prime ride times:** Encourage drivers to earn more by driving during late-night hours or to special event locations.
- **Personalized communication:** Send tailored messages with updates, tips, and encouragement to keep existing drivers motivated and support new onboarding drivers.
- **Increased demand:** Since a ride hailing business operates as a two-sided market, more customers lead to increased economic opportunities for drivers, while more drivers result in shorter wait times for passengers and likely higher satisfaction rates, creating a positive feedback loop that benefits both sides.

Next step

What can be done next from a data product's perspective?

1. Collect other data such as driver demographics, app interaction, declined/canceled rides, ride ratings, and customer service complaints to examine other possible predictors of churn or opportunities for value proposition. This comprehensive data can help better segment drivers based on a more holistic viewpoint.
2. Further investigate the high-risk segments to identify areas for intervention. The value from these drivers, if successfully retained, will outweigh the challenges of developing retention strategies.
3. Study driver behavior during the onboarding period to identify key touchpoints with the platform and company, utilizing this golden window to educate and convert them into loyal drivers. By providing comprehensive training, resources, and early support, we can ensure that new drivers feel valued and prepared, increasing their likelihood of long-term engagement.
4. Develop a churn prediction model to enable timely interventions, ensuring that drivers at risk of churn receive targeted support and incentives, such as tailored incentives, feedback loops, and proactive communication, can address specific concerns and needs, to remain engaged and productive.

Thank you!

Appendix

Data Cleaning

Steps taken to clean data and why

1. Use full outer join for ***driver_ids*** and ***ride_ids*** because there's a 10% of driver data not included in both tables.
2. Use left join for ***ride_ids*** and ***ride_timestamps*** because rides associated with non-identified drivers are irrelevant for this analysis.
3. Use date of *requested_at* as ride_date.
4. Process *arrived_at* to be equal to *picked_up_at* for ~8000 records having arrived_at time after picked_up_at time.
5. Truncate each onboarding/ride date to represent onboarding/ride week at the value of the Monday of that week.
6. Remove drivers with no onboarding date for the Cohort Analysis.

Go back

Fee Calculation

Fee Calculation Formula

Fee = (Base Fare + Cost per Mile * *ride_distance in mile* + Cost per Minute * *ride_duration in minute*) * (1 + (*ride_prime_time*/100)) + Service Fee

Assumptions on calculating fees for each ride

1. Base Fare = $2; Cost per Mile = $1.15; Cost per Minute = $0.22; Service Fee = $1.75.
2. Minimum Fare = $5; Maximum Fare = $400.
3. Service Fee is a fixed fare and is not bound to be affected by ride prime time.
4. Tax is not included and out of scope for this analysis.

Assumptions on driver's earning

1. Drivers keep 80% of the fee
2. Income tax is out of scope for this analysis.

Go back

LTV Calculation

LTV Calculation Formula

General formula: (Average ride value * Average # rides * Average Lifespan) - (Acquisition + Retention Cost).

However, in this report, the LTV is calculated as:

Average ride value ($/ride) * Average # rides (ride/driver/week) * Average lifespan (week)

Assumptions on calculating LTV

1. Average ride value is the average fare of all rides.
2. Average # ride is the average number of rides taken each week by each driver
3. Average lifespan is the average number of weeks a driver stays with the platform (based on the insight that on average, the driver retention rate by week 12/13 was around 50%, indicating that the median churn point for drivers typically occurred between weeks 12 and 13).
4. For simplification purpose, cost to acquire a driver was not included in the calculation because this cost was varied in different regions and across different timeframes.

Go back

LTV Calculation

> Multiple Regression

Features used

Feature name	Definition
avg_distance_per_drive	Average distance in meter per drive
avg_duration_per_drive	Average ride duration in second per driver
avg_arrival_duration	Average duration from accepting ride to pick up place
avg_wait_duration	Average duration from arrival to pick up to successfully pick up passengers
ride_freq_during_active	Average number of rides per active day
avg_primetime_multiplier	Average primetime multiplier across ries

Go back

LTV Calculation

> Multiple Regression

Go back

Results

OLS Regression Results

Dep. Variable:	LTV	R-squared:	0.535
Model:	OLS	Adj. R-squared:	0.532
Method:	Least Squares	F-statistic:	160.5
Date:	Thu, 04 Jul 2024	Prob (F-statistic):	1.63e-135
Time:	03:02:55	Log-Likelihood:	-7248.3
No. Observations:	844	AIC:	1.451e+04
Df Residuals:	837	BIC:	1.454e+04
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1408.5719	403.943	-3.487	0.001	-2201.433	-615.711
avg_distance_per_drive	0.1092	0.035	3.146	0.002	0.041	0.177
avg_duration_per_drive	1.3751	0.542	2.539	0.011	0.312	2.438
avg_arrival_duration	-2.5112	0.764	-3.287	0.001	-4.011	-1.012
avg_wait_duration	80.0187	11.527	6.942	0.000	57.394	102.644
ride_freq_during_active	392.0221	14.895	26.318	0.000	362.786	421.259
avg_primetime_multiplier	25.0016	7.097	3.523	0.000	11.071	38.932

Omnibus:	529.719	Durbin-Watson:	1.961
Prob(Omnibus):	0.000	Jarque-Bera (JB):	17669.024
Skew:	-2.296	Prob(JB):	0.00
Kurtosis:	24.940	Cond. No.	6.79e+04

LTV Calculation

> Multiple Regression

Interpretation

- **Model Fit:**
  - **R-squared:** 0.535 indicates that 53.5% of the variance in LTV is explained by the model.
  - **Adjusted R-squared:** 0.532 is close to R-squared, suggesting the model isn't overfitted.
  - **F-statistic:** 160.5 with a very low p-value (1.63e-135) indicates that the model is statistically significant.
- **Regression Equation**
  - LTV=−1408.57 + 0.109×avg distance per drive + 1.375×avg duration per drive − 2.511×avg arrival duration + 80.019×avg wait duration + 392.022×ride freq during active + 25.002×avg primetime multiplier

Even though the model provides some valuable insights into factors affecting driver LTV, further investigation and consideration is needed before drawing definitive conclusion:

- Focus on increasing ride frequency for drivers.
- Investigate the positive relationship between wait duration and LTV.
- Encourage drivers to take more primetime rides.
- Work on reducing arrival durations to improve LTV.
- Consider addressing potential multicollinearity in the model.

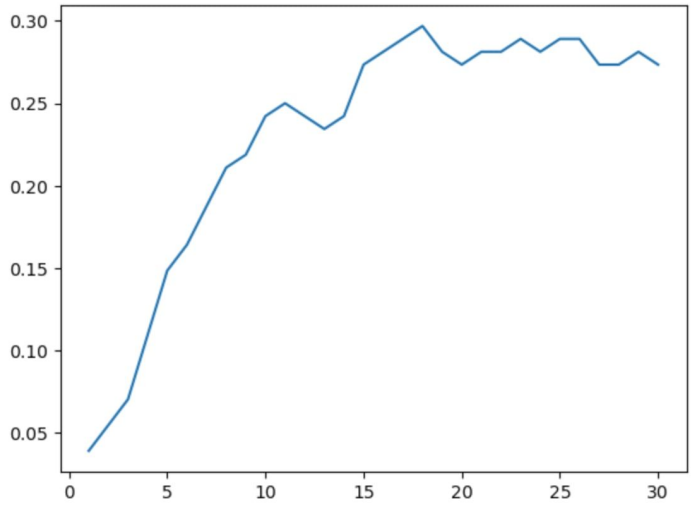
Go back

Churn Definition

How to define churn

- **Approach:** We want to find out the longest continuous inactive period where the rate of change (slope steepness) shows the smallest increase.
- **Assumptions:**
  - Use data of drivers onboarding from Week 1 to 3 only because this group had more ride data for more accurate calculation.
  - Inactive period is defined as the period after which no further activity is detected.
- **Results:** The slope at approximately 18 days shows the lowest steepness, which is a good benchmark for defining churn. For simplicity and business purposes, this period should be rounded to 3 weeks.

```
[157]: plt.plot(churn_df['l'], churn_df['churn_rate'])
[157]: [<matplotlib.lines.Line2D at 0x15ad96bd0>]
```



Go back

Hypothesis Features

Definitions of extra features

Feature name	Definition
earnings_per_active_day	Daily earning based on 80% of ride values per active days (days having rides only)
ride_freq_during_active	Number of rides taken per day during active period (outside which no further rides were observed)
avg_primetime_multiplier	Average value of PrimeTime multipliers on all rides
num_rides_in_1st_30days	Number of rides taken during the first 30 days since onboarding
avg_wait_duration	Average wait duration, calculated by averaging the difference in seconds between picked_at_up and arrived_at of all rides

Go back

Statistical Testing Results

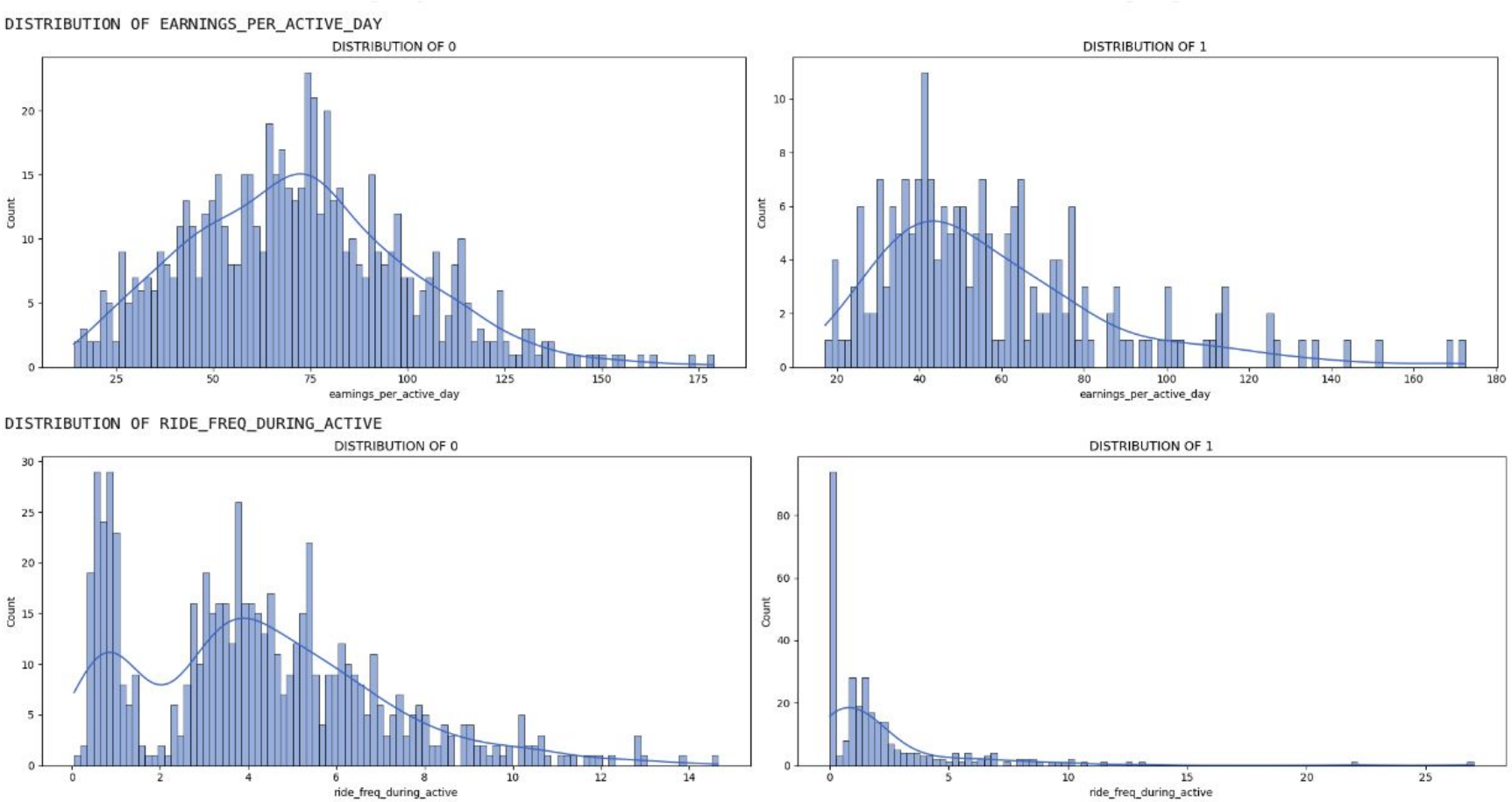
Detailed results of statistical testing on churn hypothesis

H0 and H1	Feature	Churned mean	Retained mean	t-score	p-value	Result
(H0): The mean earnings per active day is the same for churned and retained drivers. (H1): The mean earnings per active day is different between churned and retained drivers.	earnings_per_active_day	57.82	72.32	-6.2482	< 0.0005	reject the null hypothesis
(H0): The mean ride frequency during active period is the same for churned and retained drivers. (H1): The mean ride frequency during active period is different between churned and retained drivers.	ride_freq_during_active	3.25	4.24	-3.2927	0.0011	reject the null hypothesis
(H0): The mean average primetime multiplier is the same for churned and retained drivers. (H1): The mean average primetime multiplier is different between churned and retained drivers.	avg_primetime_multiplier	14.82	16.67	-2.9399	0.0036	reject the null hypothesis
(H0): The mean number of rides in the first 30 days is the same for churned and retained drivers. (H1): The mean number of rides in the first 30 days is different between churned and retained drivers.	num_rides_in_1st_30days	53.43	126.82	-13.7544	< 0.0005	reject the null hypothesis
(H0): The mean average wait duration is the same for churned and retained drivers. (H1): The mean average wait duration is different between churned and retained drivers.	avg_wait_duration	2.5	6.7	-24.0583	< 0.0005	reject the null hypothesis

Go back

Churned vs. Retained Feature Distribution >1

Detailed distributions of 5 features tested

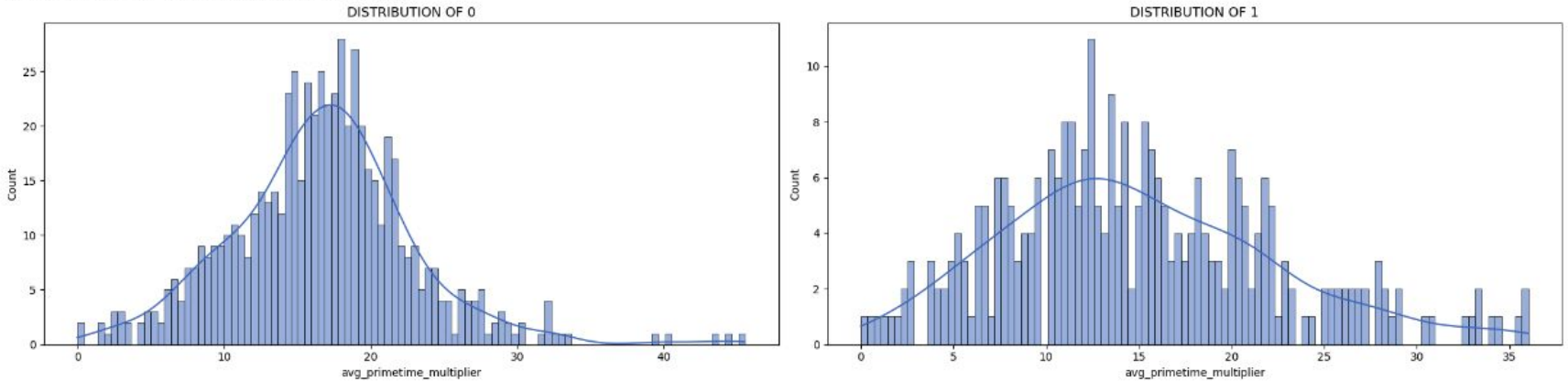


Go back

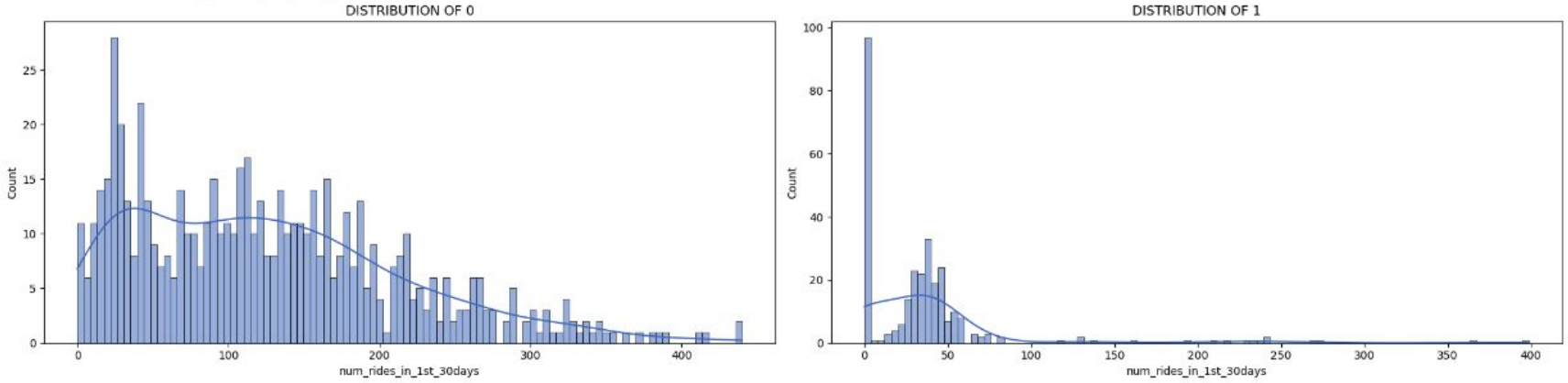
Churned vs. Retained Feature Distribution >2

Detailed distributions of 5 features tested

DISTRIBUTION OF AVG_PRIMETIME_MULTIPLIER



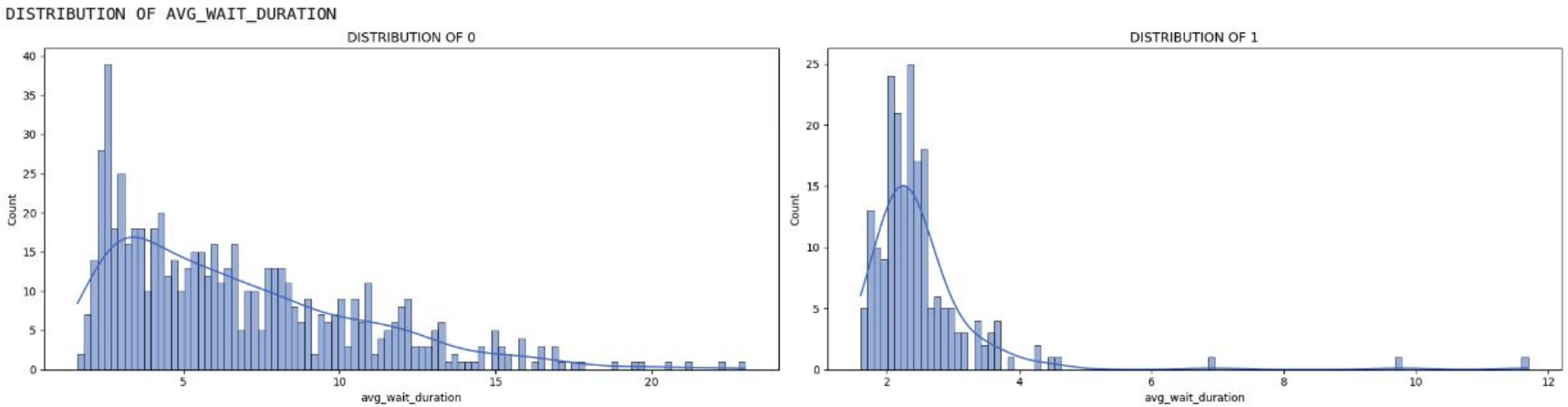
DISTRIBUTION OF NUM_RIDES_IN_1ST_30DAYS



Go back

Churned vs. Retained Feature Distribution >3

Detailed distributions of 5 features tested



Go back

RFM Features

Feature explanation

Feature name	Definition
(R) days_since_last_drive	Number of day since last drive
(F) ride_freq_during_active	Number of rides taken per day during active period (outside which no further rides were observed)
(M) earnings_per_active_day	Total daily ride values per active days (days having rides only)

Missing data handling

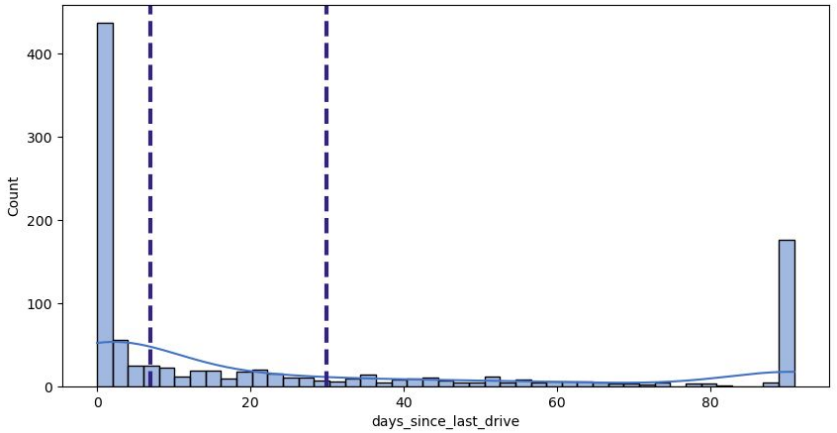
- null values of **days_since_last_drive** → 91 (total data timeframe is 90 days)
- null values of **ride_freq_during_active** → 0
- null values of **earnings_per_active_day** → 0

Go back

Appendix

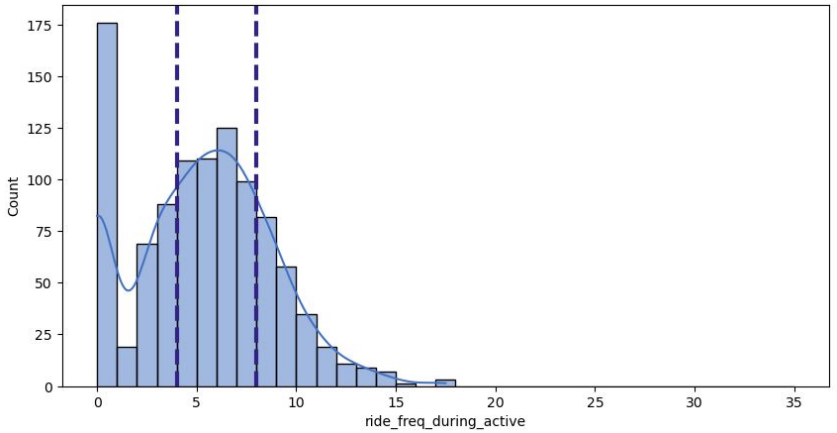
RFM Segment

Feature segmentation



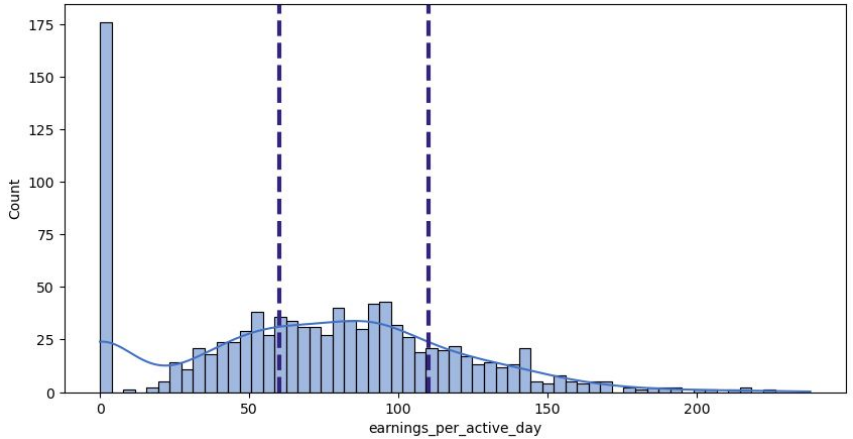
days_since_last_drive

1-Low recency: from 70 percentile up (up to 91 days)
2-Medium recency: threshold at 70 percentile (30 days)
3-High recency: threshold at 50 percentile (7 days)



ride_freq_during_active

1-Low frequency: threshold at 35 percentile (4 rides/day)
2-Medium frequency: threshold at 80 percentile (8 rides/day)
3-High frequency: from 80 percentile up (up to 17 days)



earnings_per_active_day

1-Low monetary: threshold at 40 percentile ($60/day)
2-Medium monetary: threshold at 80 percentile ($110/day)
3-High monetary: from 80 percentile up (up to $238)

Go back

RFM Segment

Sizing

RFM segments	Segment Definition	RFM Sub-group	Sizing
High performer	High value, frequent drivers	333, 332, 323	191
Loyal	Consistent drivers	322, 233, 232, 223, 222	333
At Risk	Previously engaged but showing decreased activity drivers	321, 312, 311, 221, 212, 211	176
Lost/Churn	Had not driven in more than 1 month	111, 112, 121, 122, 132, 133, 123	320

Go back

Churn Analysis

> High-risk Segment

Who are the “At Risk” drivers?

This group consists of drivers recently active (at least one ride in the last 30 days), but with decreasing ride frequency and earnings. The decline could be due to increased competition, personal issues, or dissatisfaction with platform changes. Identifying and re-engaging these high-value, at-risk drivers is crucial to prevent churn, as previous analysis shows similar individuals have left the platform without intervention.

Go back