Customer Churn Prediction and Analysis for Uber Ride-Hailing Service

I. Definition

1.1. Project Overview

Ride-hailing is an on-demand transportation service that provides an efficient travel mode by matching drivers and travelers via smartphone apps[1]. This industry is very dynamic and gives its customers lots of choices. However, customers become more knowledgeable and less patient nowadays, which easily leads them to switch to competitors. Therefore, **customer churn** becomes a very serious problem.

To meet the need of surviving in this competitive industry, the retention of existing customers becomes a huge challenge. Because retaining an existing customer is a much lower cost than acquiring a new customer[2], properly detecting riders who are likely to be churned could help improve customer retention.

This project will use transactional data from Kaggle which is about Uber trips in Peru 2010 as an experiment to predict customer churn in next 3-month threshold and analyze churner segments.

1.2. Problem Statement

Problem to be solved includes:

---

[1] Mao H, Deng X, Jiang H, Shi L, Li H, Tuo L, Shi D, Guo F. Driving safety assessment for ride-hailing drivers. Accident Analysis & Prevention. 2021 Jan 1;149:105574.

[2] Buckinx W, Van den Poel D. Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. European journal of operational research. 2005 Jul 1;164(1):252-68.

1. Identify trends in riders' activity

2. Build a prediction model with Logistic Regression to detect who is likely to churn in the next 3 months.

3. Do segmentation for churners based on RFM theory.

1.3. Metric

Evaluate performance of the binary classifiers on testing data with Accuracy Confusion Matrix, Precision, and Recall.

- Accuracy

Accuracy is a common metric to evaluate binary classifiers. It takes into account the percentage of correct predictions out of all predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$[3]

- Precison and Recall

Accuracy alone is not enough to evaluate the classifier. This is where Precision and Recall come into play.

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN}$$

Precision helps when the costs of false positives is high and Recall helps when the cost of false negatives is high.

Depending on the domain knowledge in ride-hailing service, this industry was proved to be very dynamic and give a lot of choice for riders, so customers can easily switch

---

[3] True Negative (TN): The actual negative class is predicted negative.

True Positive (TP): The actual positive class is predicted positive.

False Negative (FN): The actual class is positive but predicted class is negative.

False Positive (FP): The actual class is negative but predicted class is Postive.

to competitors. The cost of wrong non-churner prediction (false negatives) will be larger than the cost of wrong churner prediction (false positives), so Recall will be more mattered than Precison.

II. Analysis

2.1. Data Exploration

Date span of the transactional data is from 1/1/2010 to 12/31/2010. There are 23,111 instances (number of trips) with 28 attributes in total. This project only uses 8 attributes as shown in Table 1. In addition, there are 1390 riders and 168 drivers in the dataset.

Table 1. Data information

Column	Data Type	Null count	Description
journey_id	object	0	Unique identifier for each ride
user_id	object	0	Unique identifier for each rider
start_at	datetime	0	Datetime when a ride is booked on Uber app
end_state	object	12	End state of each ride (*drop off*, *driver cancel*, *rider cancel*, *no show*, *not found*, *failure*).
price	float64	398	Price for each trip (some price can be considered cancellation fee).
rider_score	float64	7721	Score of riders rated by drivers for each trip.
distance	float64	263	Distance from pickup point to drop off point.
duration	float64	263	Duration from pickup point to drop off point.

	price	distance	duration	rider_score
count	22713.000	22848.000	22848.000	15390.000
mean	2752.739	10883.953	638.831	4.755
std	3025.390	202573.499	1788.661	0.841
min	0.000	0.000	0.000	0.000
25%	1700.000	0.000	0.000	5.000
50%	1911.000	4660.000	218.000	5.000
75%	3597.000	9290.250	667.000	5.000
max	55974.000	14037219.000	83807.000	5.000

Figure 2. Descriptive Statistics

From Figure 2, we can find that lots of outliers are presented in three columns: price, distance, duration. We will tackle with this problem in later sections.

2.2. Data visualization

The plots show below is to identify any trend in riders' activity. This is helpful for dealing missing value and get insight to extract features for prediction model.

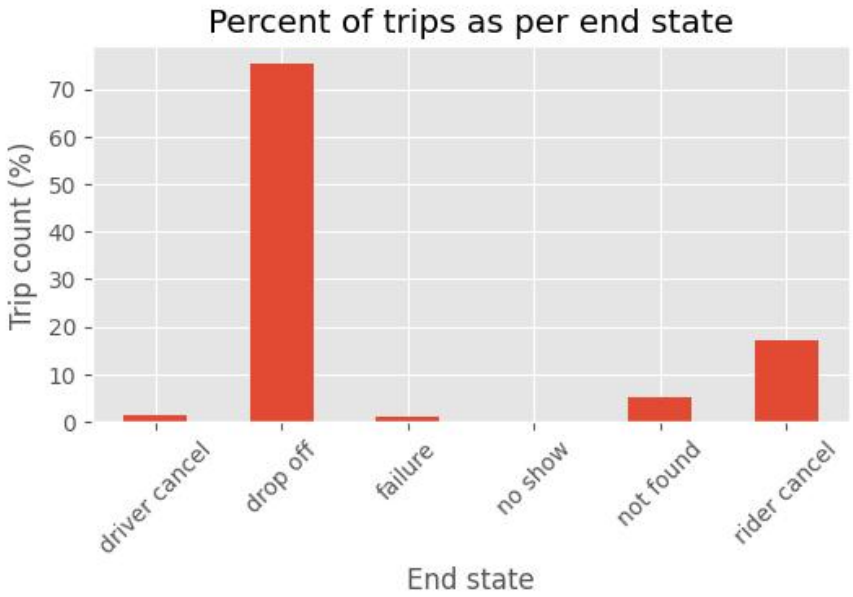


Figure 3. How were status of the trips?

Figure 3 shows the distribution of end state of all trips. Over 70% are drop off which means riders were arrived at drop off point successfully. The rest are unsuccessful trips with several reasons, the most common reason is that rider cancel the booking.

In addition, out all the trips that were canceled by rider, there are about 5% of trips being charged cancellation fee.

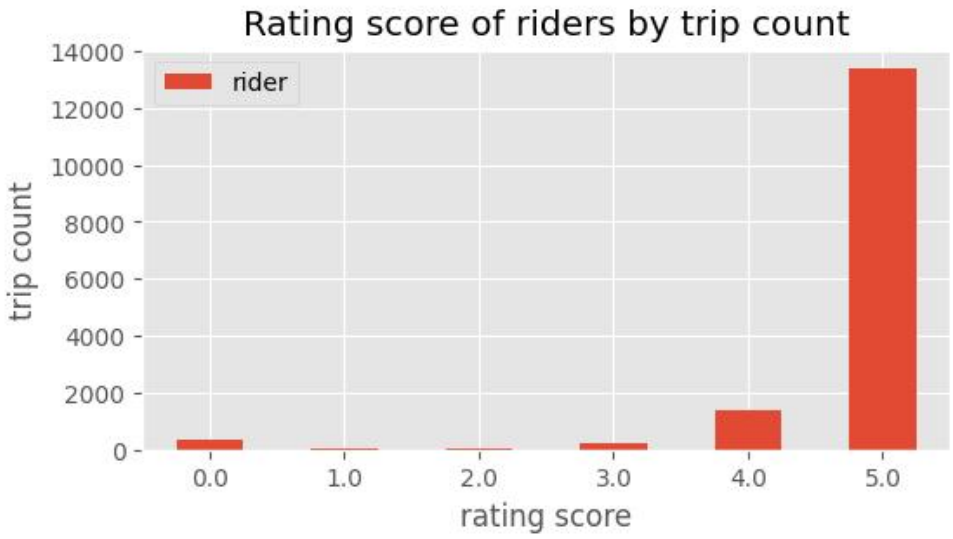


Figure 4. Number of trips were given score in the scale 0-5 for rider

Figure 4 shows that riders typically given score 5 and then 4. Because driver can decide to and NOT to give score for rider after each trips, this is the reason why the column “rider_score” hold the most missing value (7721).

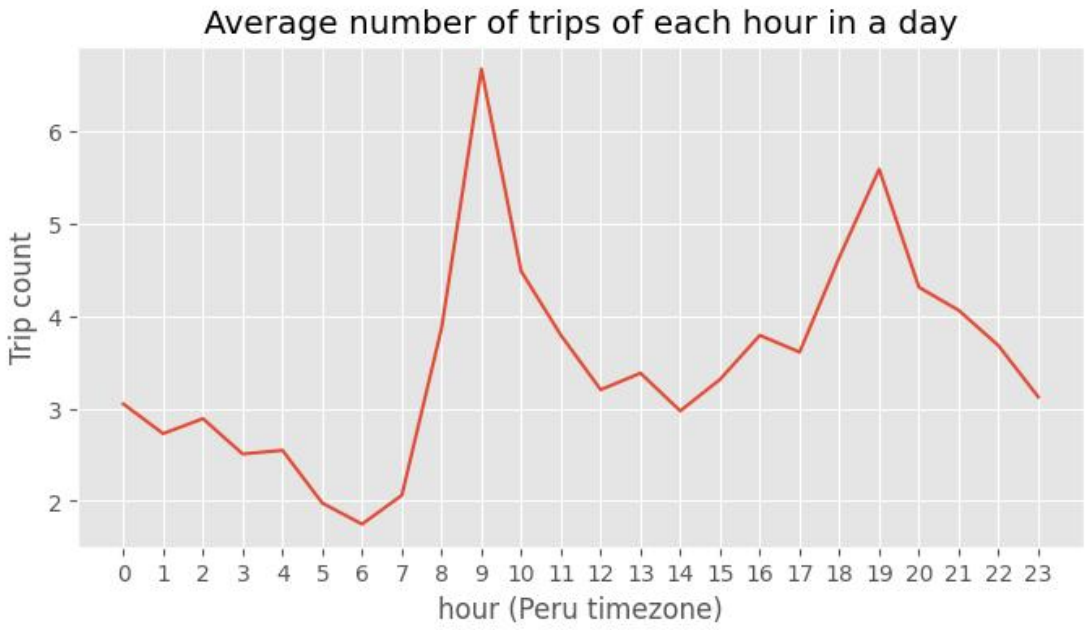


Figure 5. When was the peak period in a day?

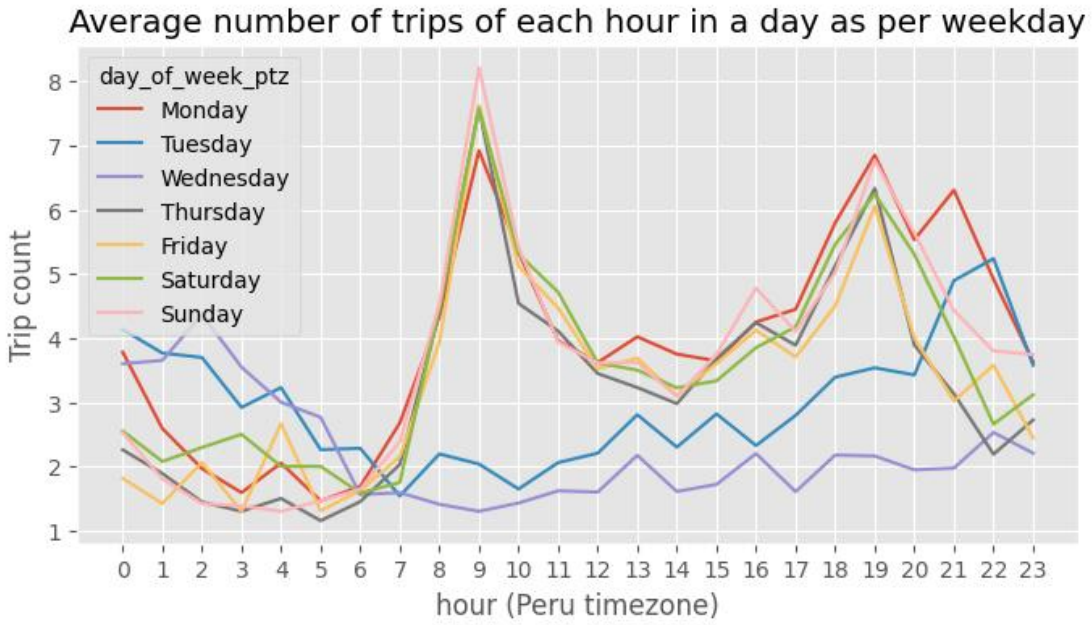


Figure 6. When was the peak period in a day (by day of week)?

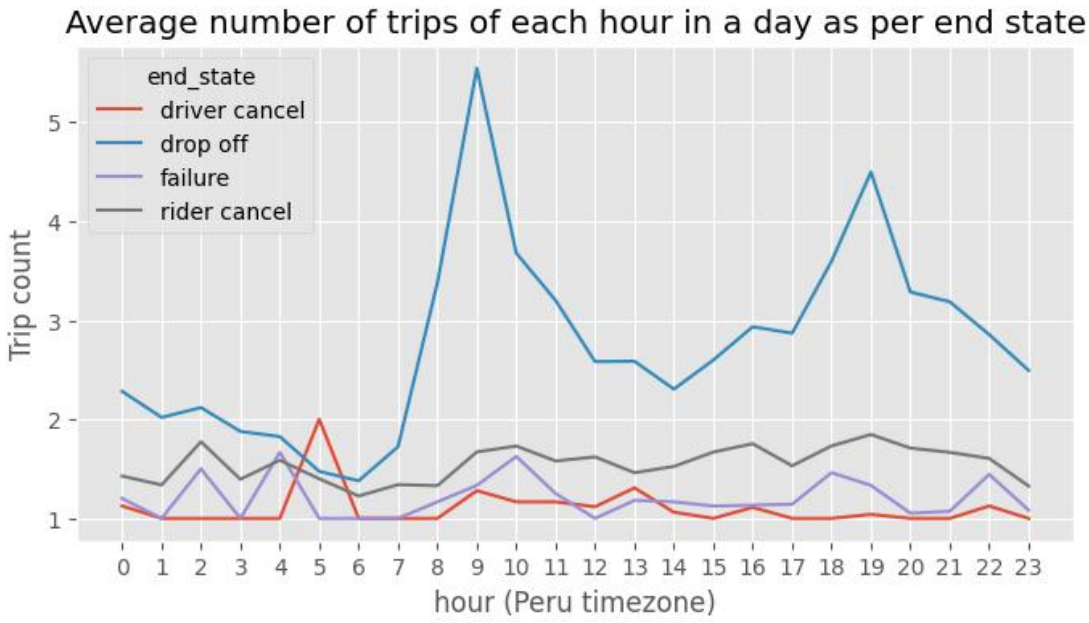


Figure 7. When was the peak period in a day (by end_state)?

Figure 5, 6, 7 show that over the day of week or end state, the peak period in a day is from 9-10am and after 5pm.

III. Methodology

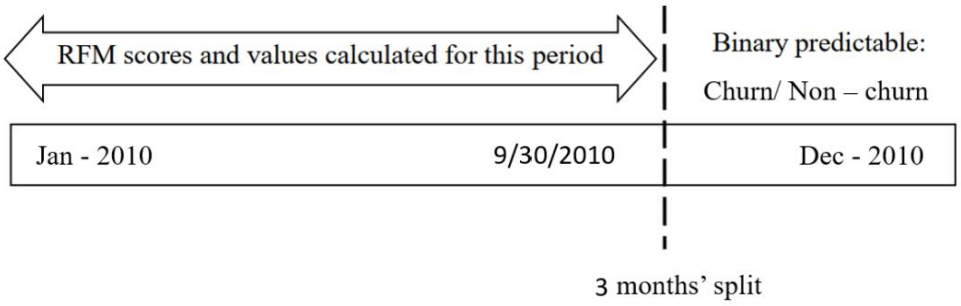
3.1. Data preprocessing

The preprocessing step done in "Feature engineering" notebook mainly relies on RFM theory[4]:

- R - Recency: How recent was the customer's last behavior?
- F - Frequency: How often did this customer perform that behavior in a given period?
- M - Monetary: How much money did the customer spend in a given period?

[4] Aleksandrova, Y., 2018. Application of machine learning for churn prediction based on transactional data (RFM analysis). In 18 International Multidisciplinary Scientific Geoconference SGEM 2018: Conference Proceedings (Vol. 18, No. 2.1, pp. 125-132).

And data labeling method from Xia and He (2018)[5] as below.



(Source: author)

Figure 8. Feature Engineering Idea

The preprocessing and transformation process consists of the following steps:

1. **Divide data**

Divide the transactional data into observation period (from 2010-01-01 to 2010-09-30) and forecast period (from 2010-10-01 to 2010-12-31). Observation data will be used for feature engineering and forecast data will be used for data labeling.

2. Handle missing value

Four columns need missing value handling: end_state, price, distance, duration. Because the dataset is quite small for modeling, compute missing value as many as possible.

3. Transform data

Aggregate the transactional data at user level and extract features based on RFM theory.

4. **Label data**

[5] Xia, G. and He, Q., 2018, March. The Research of online shopping customer churn prediction based on integrated learning. In *Proceedings of the 2018 International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2018), Qingdao, China* (pp. 30-31).

- If the riders did not have any `drop off` trip during the forecast period, such riders are considered to be churn within `3-month threshold`, marked as `1`.

- If the rider has at least one `drop off` trip during the forecast period, such riders are NOT considered to be churn, marked as `0`.

5. **Handle missing value for features from step 3&4**

Before passing the data into machine learning model, we will get columns with null value filled up. Because the dataset size is quite small, we take a very naive approach to fill NA with zero.

3.2. Implementation

The implementation process can be split into two main stages:

3.2.1 The classifier training stage

Build a binary classifier to predict customer churn with Logistic Regression done in the notebook “Build classifier”. Logistic Regression are not much impacted by the presence of outliers because the sigmoid function tapers the outliers. Therefore, outliers will be keep in the data as they are and only revise once they really have a bad impact on the model performance.

3.2.2. Churn segmentation

This stage was done in the notebook “Churn segmentation” based on the idea of Xia and He (2018) and RFM theory.

At the same time, regardless of the impact of the time node division on the customer's recent behavior date, divide the two variables F and M into two states:

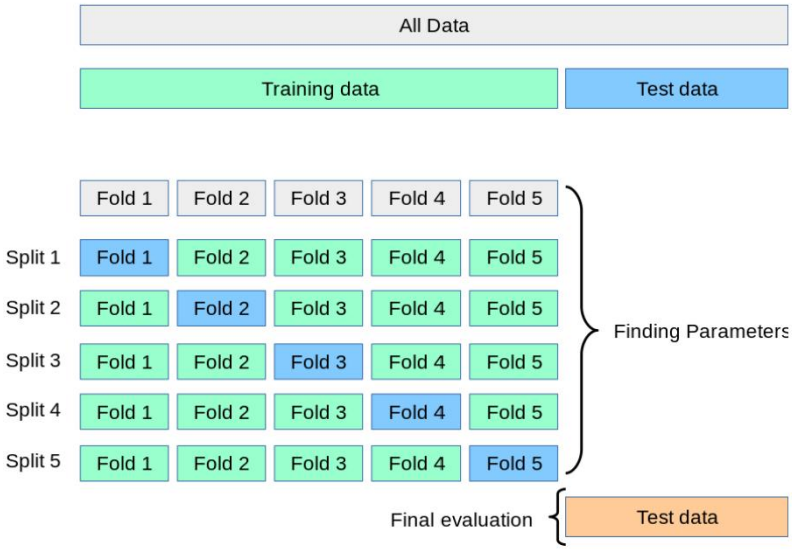
- F1 is lower than the average frequency, F2 is higher than the average frequency

- M1 is lower than the average spending, M2 is higher than the average spending.

Churn value can be divided into 4 group: F_1M_1, F_1M_2, F_2M_1, F_2M_2.

3.3. Refinement

Data was split into training set (80%) and testing set (20%). Then doing 10 fold cross validation on training data to see how the model performs with accuracy, precision, recall before generalized on testing data. 10 fold cross validation means that one group is tested using the classifier trained on the remaining nine groups sequentially.



(Source: scikit-learn.org)

Figure 9. K-fold cross validation

Doing 4 trials with 10 fold cross validation to obtain the best classifier are as below:

- Trial 1: Leave everything as they are for cross validation
- Trial 2: Apply ANOVA f-test[6] on training set to rank feature importance, find top k important feature that were the most useful at predicting the target variable. Then do cross validation with these k features.

[6] Kumar M, Rath NK, Swain A, Rath SK. Feature selection and classification of microarray data using MapReduce based ANOVA and K-nearest neighbor. Procedia Computer Science. 2015 Jan 1;54:301-10.

- Trial 3: Normalize feature with zscore due to skewed distribution of the data and then leave everything as they are for cross validation.
- Trial 4: Use normalized training data from Trial 3 and feed top k important features from Trial 2 to cross validation.

IV. Results

4.1. The classifier training stage

4.1.1. Model Evaluation and Validation

Use Logistic Regression model from sklearn with solver='liblinear' and compare different average accuracy of 10 fold cross validation with 4 trials. Then take model with highest average accuracy and highest recall for being generalized on the test data.

4.1.2. Justification

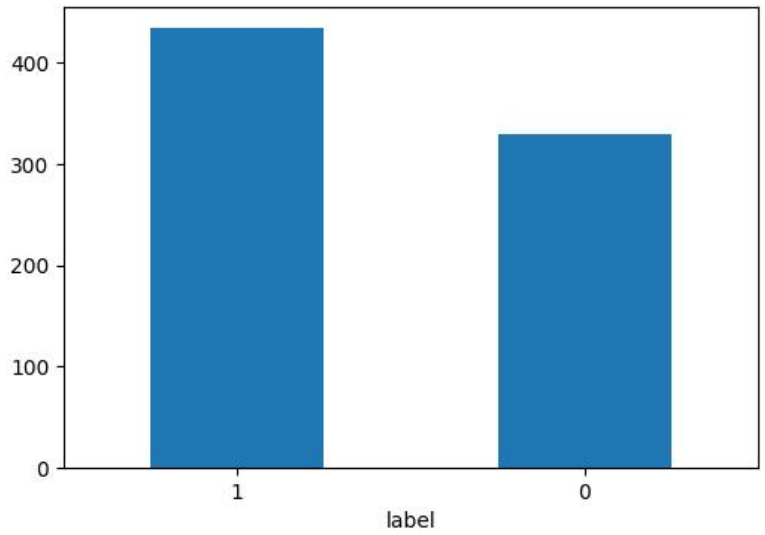


Figure 10. Class Distribution

Figure 10 show that the number of class 0 and 1 are quite evenly distributed so no need to deal with class imbalance.

Table 2. Evaluate Model Performance

	10 fold cross validation on training data		**Prediction result on testing data**		
	Average accuracy (%)	Standard deviation	Accuracy (%)	Precision (%)	Recall (%)
Trial 1	76.84	3.58	72.92	72.46	87.72
Trial 2	79.06	4.78	73.44	71.94	89.29
Trial 3	78.54	5.42	74.48	76.86	81.56
Trial 4	80.37	4.2	75.52	77.69	82.46

Table 2 shows that over the four trials, accuracy on the test data is ranged from 72% to 75%. We can clearly find the trial 4 has the highest accuracy but accuracy alone is not enough to conclude that the model from trial 4 is the best.

As explained in the *Metric* section, the cost of wrong non-churner prediction (false negatives) will be larger than the cost of wrong churner prediction (false positives), so Recall will be more mattered than Precison. Recall of trial 2 is significantly higher than trial 3 and 4 although the accuracy of these two trials are higher than trial 2's accuracy.

With the highest recall of 89.29% and an acceptable accuracy of 73.44%, the model from trial 2 is determined as the best model to predict customer churn in the next 3-months. Now let's see how the process of trial 2 was done:

1. Apply ANOVA f-test on training set to rank feature importance. Figure 11 shows the result.

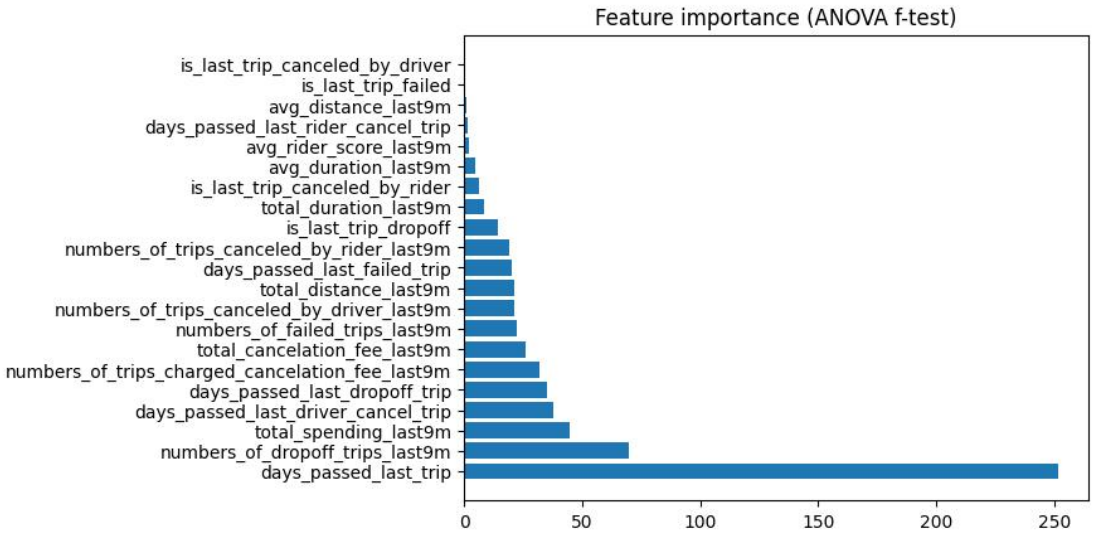


Figure 11. ANOVA f_test for trial 2

2. See what is the average accuracy over the k important feature for cross validation.

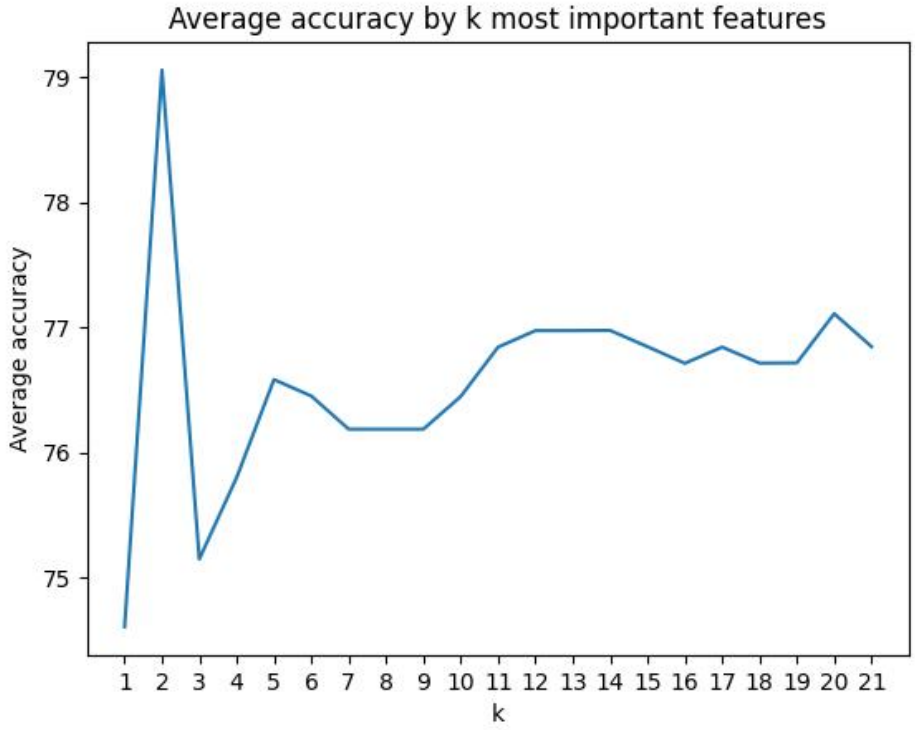


Figure 12. Accuracy over k important feature for cross validation

Although k=2 yields the highest accuracy but k=2 seems not to be practical. k=14 were used for the classifier instead.

3. With top 14 important feature selected above, generalize the Logistic Regression model on testing data, get the following metric.

	precision	recall	f1-score	support
0	0.76	0.50	0.60	78
1	0.72	0.89	0.80	114
accuracy			0.73	192
macro avg	0.74	0.70	0.70	192
weighted avg	0.74	0.73	0.72	192

Accuracy: 73.44

Figure 13. Classification report

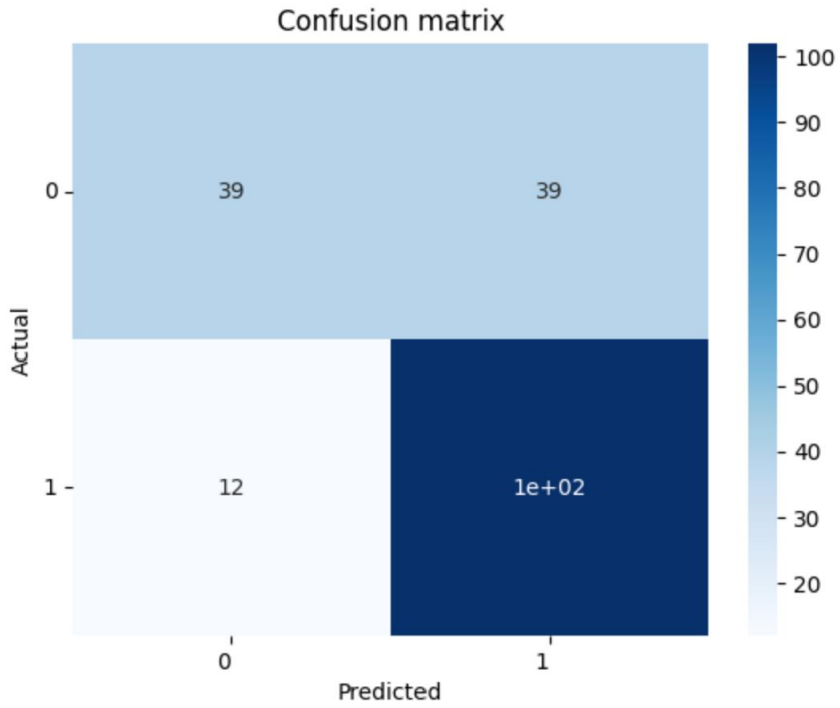


Figure 14. Confusion matrix

4.2. Churn segmentation

	Churn_Segment	Number_of_riders	Rate(%)
0	F2M2	460	48.117155
1	F1M1	376	39.330544
2	F1M2	113	11.820084
3	F2M1	7	0.732218

Figure 15. Confusion matrix

**F1: lower than the average frequency*

**F2: higher than the average frequency*

**M1: lower than the average spending*

**M2: higher than the average spending*

Figure 15 shows that F2M2 presents the largest proportion of 48,12 percent among churners. Such proportion is a little bit strange because the higher the value of Frequency the more loyal are the customers of the company, similar for Monetary[7]. It could be due to the presence of extreme values and beyond the scope of this project. In general, the likelihood of churn tendency is significantly reduced as the frequency (F) and the spending (M) increase, since the rates of churn segments descend from F1M1 (39,33%), F1M2 (11,82%) to F2M1 (0,73%).

V. Conclusion

5.1. Reflection

The project provides an important guide for on-demand mobility companies to improve customer adhesiveness. Although the scope of this project was narrowed in

[7] Wang CH. Apply robust segmentation to the service industry using kernel induced fuzzy clustering techniques. Expert systems with applications. 2010 Dec 1;37(12):8395-400.

the 4-wheel ride-hailing industry, the managerial implications can be widely applied for many types of ride-hailing services (e.g., two-wheel, four-wheel) since the characteristics are almost the same.

In the context of ride-hailing services, the relationship is non-contractual so it is quite difficult to judge whether or not customers will give up the enterprise completely in the non-contractual scenario. This project studies intermittently loss of the riders within a time threshold of 3 months instead, but the riders may make transactions beyond such time threshold. Therefore, it is desirable to analyze and understand churn segments for better retention solutions in the ride-hailing industry.

The result shows that the likelihood of churn is significantly reduced as the frequency and the spending increase. For retention solutions, managers should start from Recency value first and then Monetary, try to push rider's activity more frequently and the spending as much as possible, this will effectively reduce churn rate in a dynamic and competitive business environment. Although low-value customers (F1M1 churn segment) are less valuable to the business and can be ignored from the traditional management concept, such customers should not be ignored and should be considered as a crucial challenge that mobility companies must be solved.

5.2. Improvement

The model accuracy now is just acceptable and has not yet reached 80%. We can try to use ensemble method or hybrid method to combine models instead of a single classifier. These methods can take advantage of the various prediction models to increase the reliability and stability of the prediction results.

In addition, study interesting outliers by reporting findings with and without outliers since this method can explain any difference in the results[8].

[8] Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intelligent data analysis. 2002 Jan 1;6(5):429-49.