VPBank

Case study of using Ride hailing to improve decision making for Retail customers

Dmytro Kolechko, PhD, VPBank CRO

July 2023

Executive summary

1 Imperatives

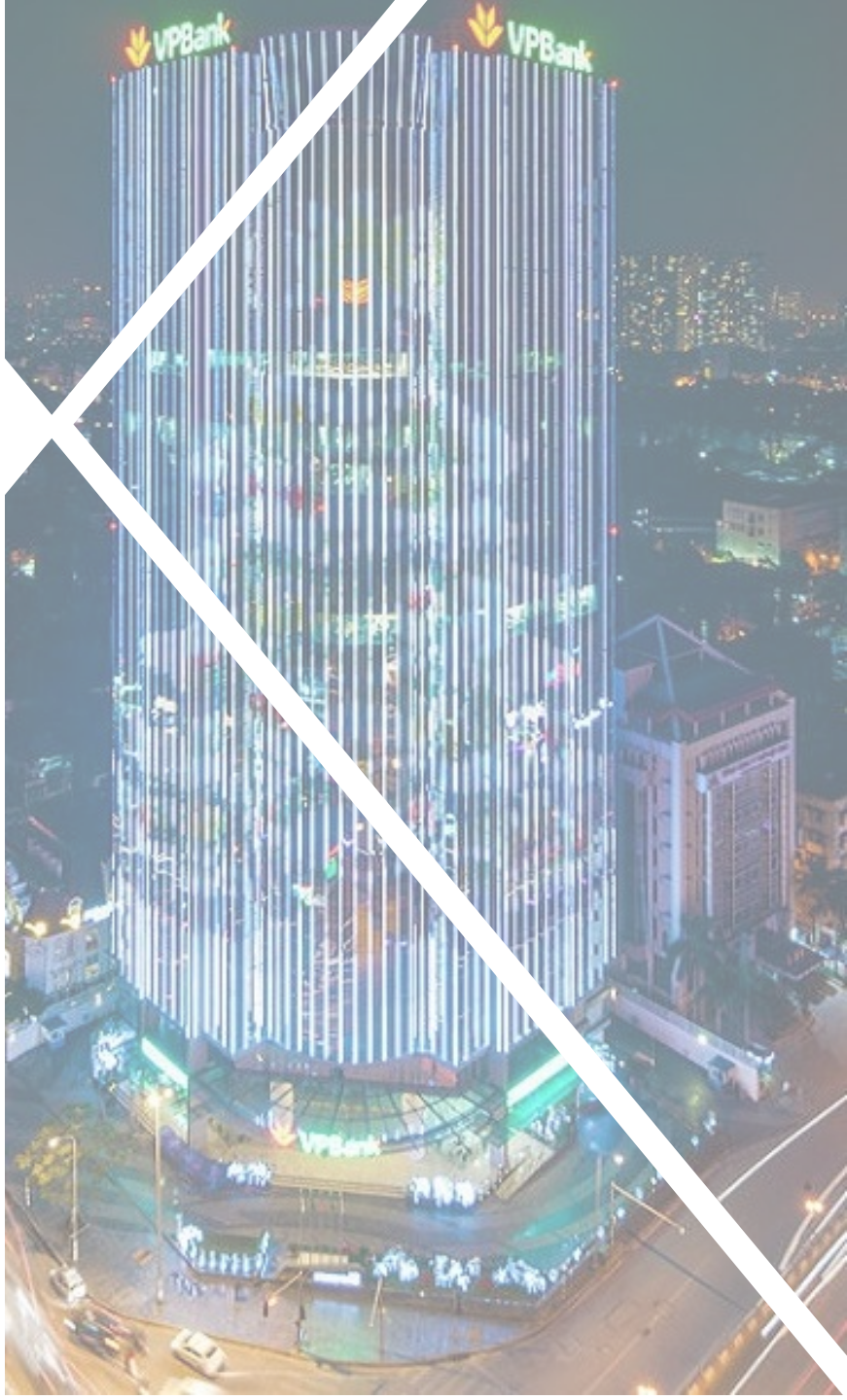
- Vietnam ride hailing market size grows rapidly with 22.7% compound annual grow rate for the period from 2018 to 2027.
- Develop models using ride hailing data to a) predict credit worthiness of customers and b) compare model productiveness with traditional model and telco model

2 Scope of work

- Develop **3 models** for selling **Credit card** and **Unsecured personal loan (UPL)**
- **Compare Gini** of models with **traditional model** and **Telco model**.

3 Results

1) Three models show **good discrimination power**.
2) Ride hailing data has **higher predictive power** than traditional data but weaker than Telco data.
3) Since launching the Credit card product for ride hailing customer in September 2022, **45K** credit cards were issued. It is needed to build Fraud trigger/Fraud scorecards to maintain the proper quality of portfolio.



1. Overview

Literature review

- Since 2018, Grab has cooperated with different partners to introduce micro-loan for Grab drivers and businesses that use its GrabPay services in South East Asia in which the drive's style could pay a part in assessing whether they receive a loan (according to Head of Grab pay speech).
- Uber has also announced about car loan for Uber drivers since 2013 but it seems purely business partnership with auto dealer/financing provider for payment, not using driver behavior for underwriting.
- While a lot of research have mentioned alternative data source used for credit scoring such as telco, consumer behavior, psychology, rare research mentions about ride hailing data used for issuing credit to customer. Our research shows that a model using ride-hailing data can be effectively implemented in credit processes for retail products such as Credit Cards, Unsecured Personal Loans, etc. while keeping credit risk under control.

Aims of the research

Research focus on:

- ✓ Develop models using ride hailing data to acquire customers from one of the biggest ride hailing platforms in Vietnam.
- ✓ Compares the outcome of algorithms such as Logistic Regression, Gradient Boosting Trees and Recurrent neural networks,
- ✓ Compares predictive power of ride hailing data with traditional model and Telco data.
- ✓ Lessons learned during scoring models development.
- ✓ Results of the practical implementation of ride hailing credit scoring models in production.

2. Data overview

2.1. Data categories

Demographic data

Application form
Biographic application

- Age, gender
- Registration time, 1st transport, 1st delivery
- Vehicle, phone, documents,..

Trip's information

Rating, ride time, waiting time
Trip performance metrics

- Trip performance metrics: frequency, time, region, fare and their combination
- Driver rating, user rating
- Ride time, waiting time

Services information

Top-up, Airport, Ticket, Reward, …
Redeemed points

- Number and amount of transactions for all services
- Total point redeemed, number of times getting redeemed points
- Failed payment recharged

Credit historical information

Historical behavior of customer at Financial institutions

- Credit relation at VPBank: product, limit, exposure
- Debt group at VPBank and other Financial institution
- Number of inquiries from financial institutions

2. Data overview

2.2. Trip performance metric

Performance metric of the trip is **evaluated** on **4 aspects**

Frequency

Analysis of the frequency of trips based on:

- ✓ Requested, accepted, cancelled, completed
- ✓ Average of user rating, driver rating
- ✓ Distance of trip

Time

Analysis by hour and day of the week of the trip:

- ✓ Rush hour (6am – 9am and 4pm – 5pm)
- ✓ Night shift (11pm – 5am)
- ✓ Weekdays (Monday to Friday)
- ✓ Weekend (Saturday and Sunday)

Region

Division of pick-up and drop-off areas by locations based on digital map:

- ✓ Urban area, suburban area, center
- ✓ Area with commercial centers, shopping mall, travel destination
- ✓ Areas with high transportation demand: airports, stations, ...

Fare

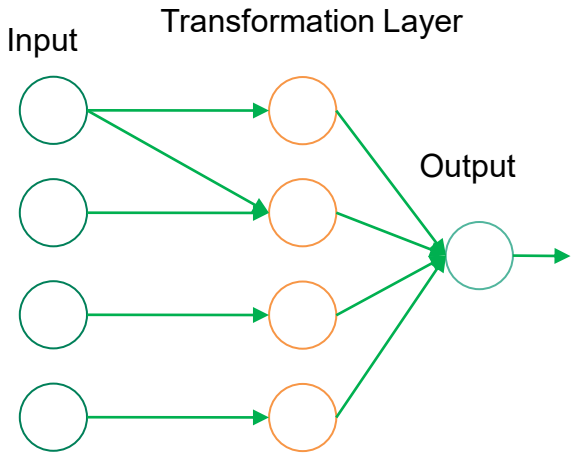
Analysis by the amount calculated by trips:

- ✓ Fare trips, tip amount (or income/ incentive of drivers)
- ✓ Promotions: discount, redeemed point, …
- ✓ Payment type: cash, non-cash

3. Model development

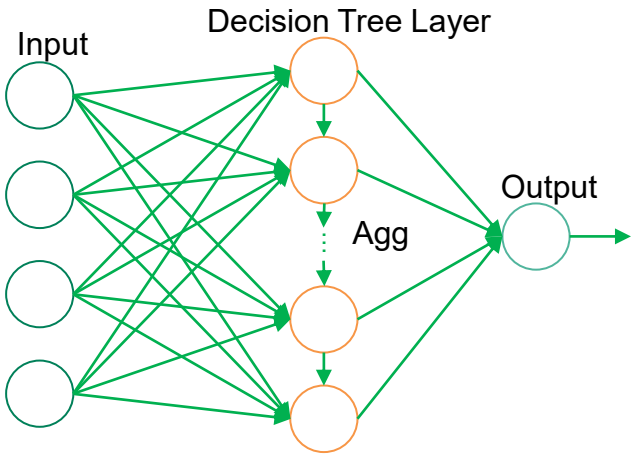
3.1. Model approaches

Logistic Regression



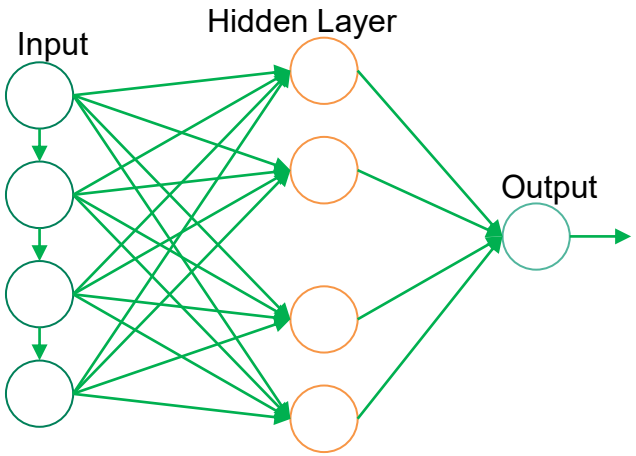
- Used as the traditional model of banking for decades.
- Easy to explain impact of variables to the output.
- Rely heavily on variable creation.

Gradient Boosting Tree



- New algorithm in the last decade
- Can find “deep” relationships between the variables.
- The identification of cut-off points in trees is completely automatic.
- Not easy to explain impact of variable to the output.
- Rely heavily on variable creation.

Neural Network



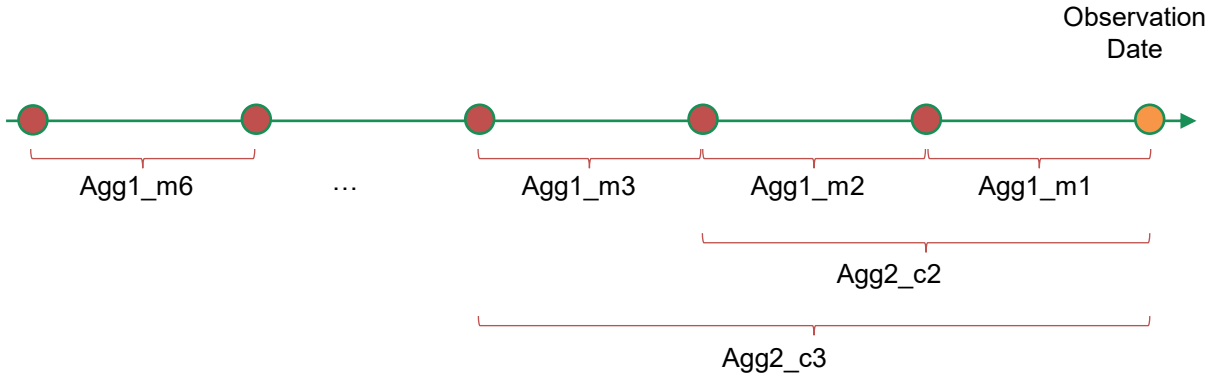
- Play an important role in deep learning algorithms.
- Can handle time-series data (recurrent neural network).
- Not easy to explain impact of variable to the output.
- Not rely heavily on variable creation

3. Model development

3.2. Data preparation for Logistic Regression and Gradient Boosting Decision Tree (GBDT)

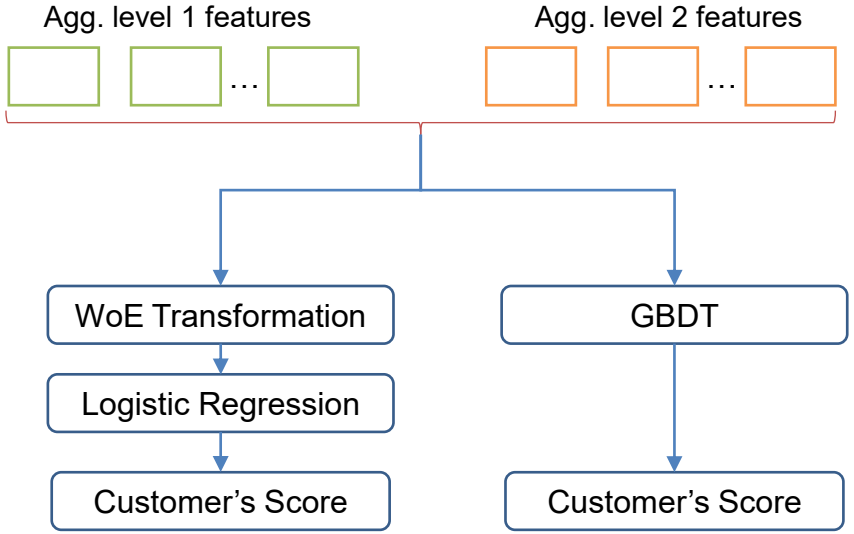
Feature engineering

- Sample: booking information within 6 months prior
- Variable creation: aggregation by single month (called agg. level 1) or accumulated over months (called agg. level 2).



Agg Level	Type	N	Example
Agg. level 1	Aggregated function	210	• Sum_fare, max_distance, … • Number_trip, number_night_trip,… • Number_trip_to_airport, …
	Ratio	180	• Percentage_of_night_trip,… • Percentage_of_weekend_trip,…
Agg. level 2	Aggregated function	600	• Max_fare, Avg_fare
	Ratio	420	• Percentage_of_night_trip,… • Percentage_of_weekend_trip,…

Model architecture

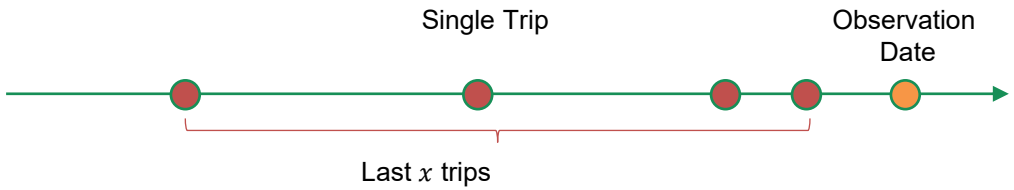


3. Model development

3.3. Data preparation for Recurrent Neural Network (RNN)

Feature Engineering

- The last x trips within 6 months prior to the customer's observation date were used to build the predictive model (in which x is average of number of trips during 6 months prior to the observation date)
- For customers who do not have enough x trips, the padding method is used



The following variables are created from the available data:

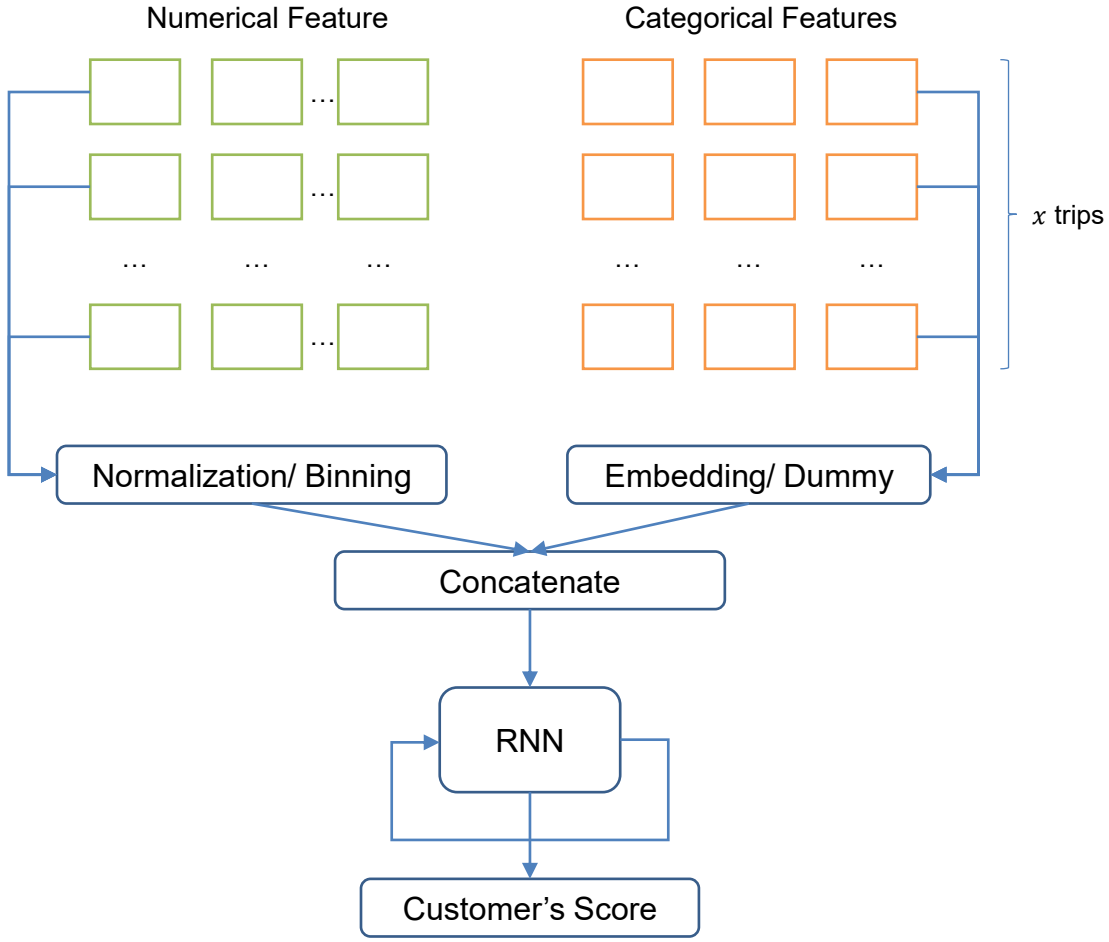
- Distance (in days) between booking date and observation date.
- The categorical variables represent the component of the booking date (booking time, day of the week).

We use **Single Trip** feature to train RNN model

Type	N	Features
Numerical	3	Amount, distance, day to observation,
Categorical	4	Drop Type, Pickup Type
Categorical	2	Hour of trip, Weekday of Trip

Model architecture (RNN)

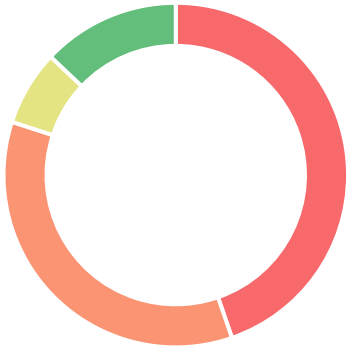
We design recurrent neural network model with the following structure:



4. Variable Assessment

4.1. Demographic information

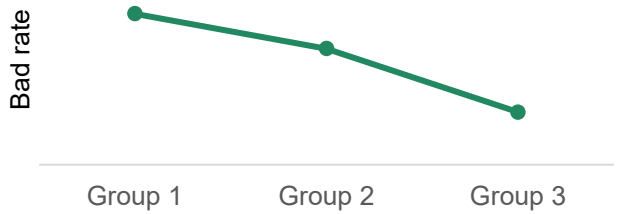
Number of features: 85	
Un-predictive	38
Weak	30
Medium	6
Strong	11



Assessment: Medium predictive power

IV	Assessment
< 0.02	Un-predictive
[0.02; 0.1)	Weak
[0.1; 0.3)	Medium
>= 0.3	Strong

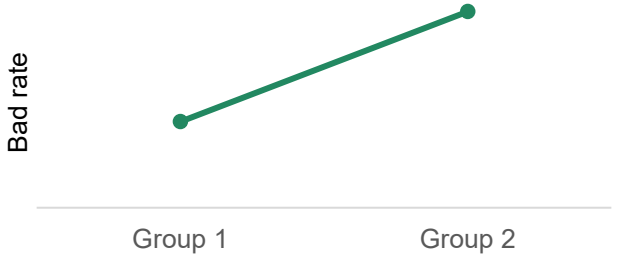
Classes of driver licenses



Expected trend: ↓ Actual trend: ↓

The higher classes of driver licenses are better

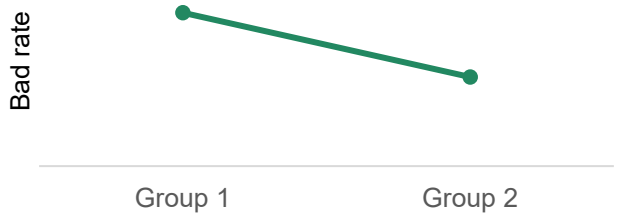
Phone



Expected trend: ↑ Actual trend: ↑

Customers who use luxury phones are better

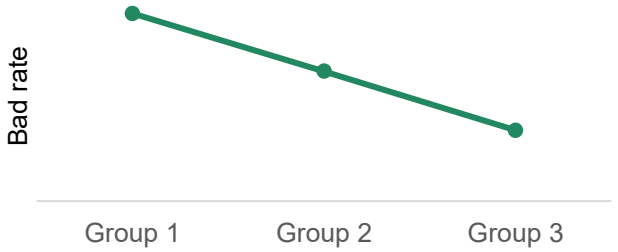
Time from first app installation



Expected trend: ↓ Actual trend: ↓

The longer time is better

Time since registration date



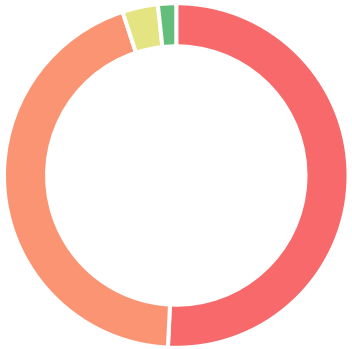
Expected trend: ↓ Actual trend: ↓

The longer time is better

4. Variable Assessment

4.2. Trip's information

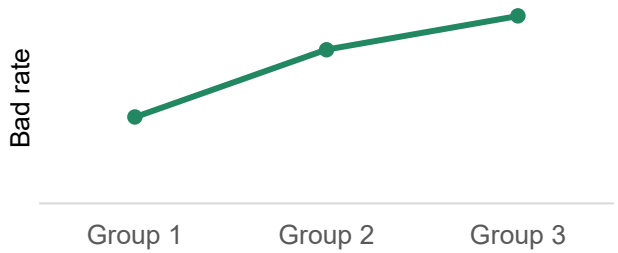
Number of features: 1229	
Un-predictive	624
Weak	543
Medium	41
Strong	21



Assessment: Medium predictive power

IV	Assessment
< 0.02	Un-predictive
[0.02; 0.1)	Weak
[0.1; 0.3)	Medium
>= 0.3	Strong

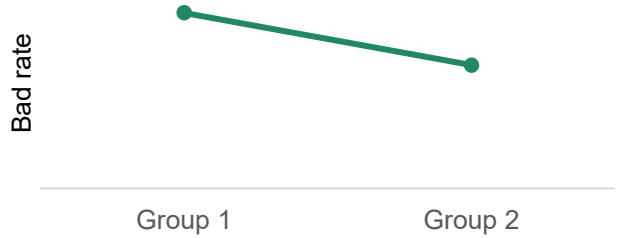
Number of complete trips paid by cash during last 3 months



Expected trend: ↑ Actual trend: ↑

Cash payment is worse than non-cash payment.

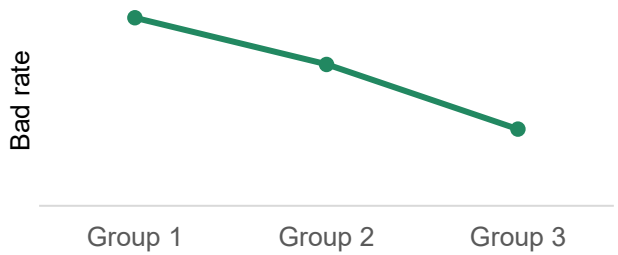
Total fare amount of complete trip during last 3 months



Expected trend: ↓ Actual trend: ↓

The higher consumer is, the better customer is

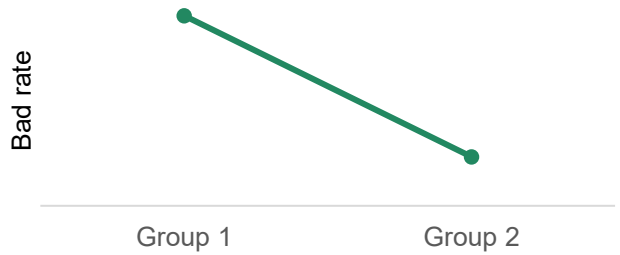
Average balance on driver's wallet divided by average income in last 3 months



Expected trend: ↓ Actual trend: ↓

The higher the ratio is, the better the customer is

Number of trips from/ to airport in last 6 months



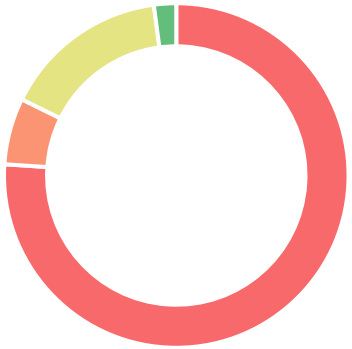
Expected trend: ↓ Actual trend: ↓

The more trips customer uses, the better customer is

4. Variable Assessment

4.3. Services information

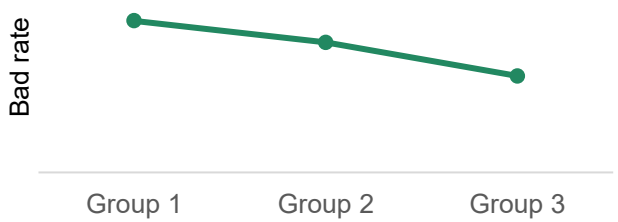
Number of features: 96	
Un-predictive	73
Weak	6
Medium	15
Strong	2



Assessment: Medium predictive power

IV	Assessment
< 0.02	Un-predictive
[0.02; 0.1)	Weak
[0.1; 0.3)	Medium
>= 0.3	Strong

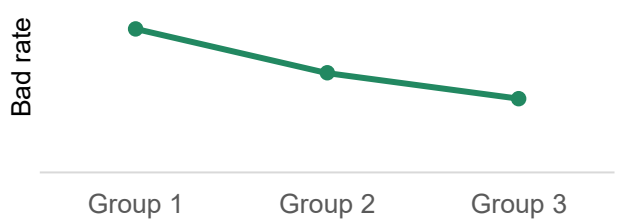
Customer rating (defined by ride hailing company)



Expected trend: ↓ Actual trend: ↓

The higher grade customer are better

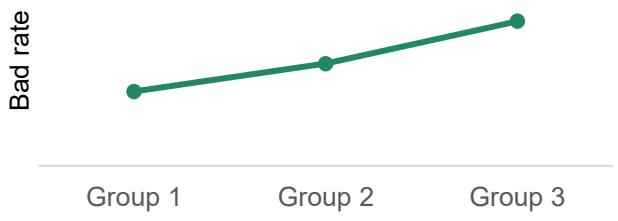
Number of times using ride hailing service in last 6 months



Expected trend: ↓ Actual trend: ↓

The more services customer uses, the better customer is

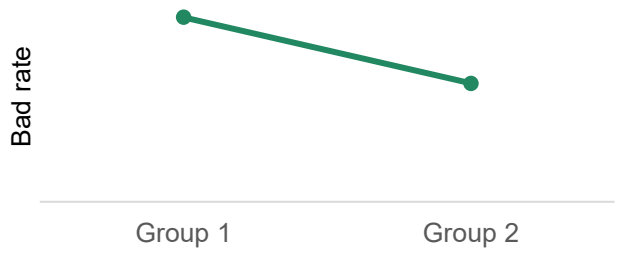
% Cash payment for ride hailing services in last 6 months



Expected trend: ↑ Actual trend: ↑

Cash payment is worse than non-cash payment

Total payment for ride hailing services in last 6 months



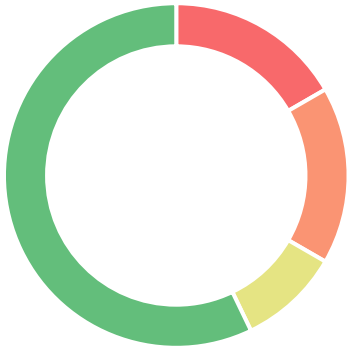
Expected trend: ↓ Actual trend: ↓

The more services customer uses, the better customer is

4. Variable Assessment

4.4. Credit historical information

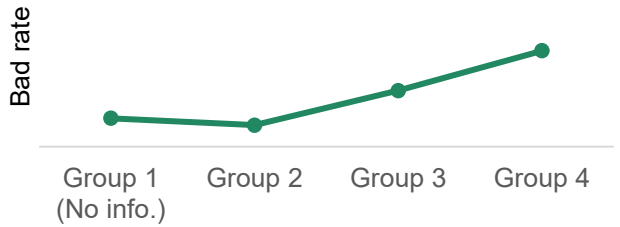
Number of features: 42	
Un-predictive	7
Weak	7
Medium	4
Strong	24



Assessment: Medium predictive power

IV	Assessment
< 0.02	Un-predictive
[0.02; 0.1)	Weak
[0.1; 0.3)	Medium
>= 0.3	Strong

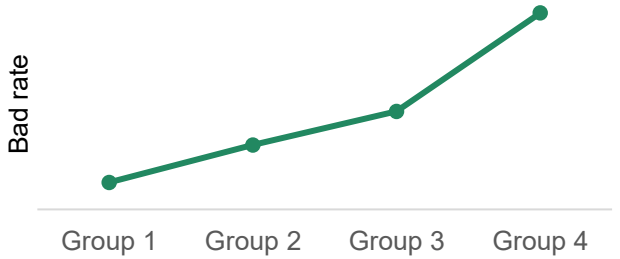
Total outstanding balance of loan



Expected trend: ⬆ Actual trend: ⬆

Customers with higher outstanding balance may face more financial difficulties than other customers

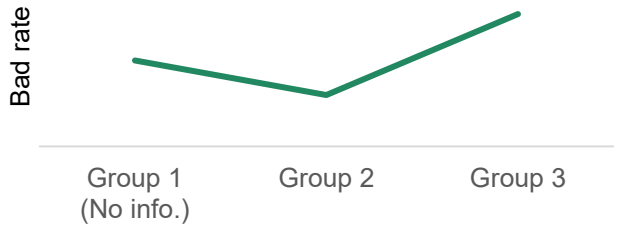
Number of inquiries from financial institutions in last 3 months



Expected trend: ⬆ Actual trend: ⬆

Customers with more inquiries from financial institutions have higher bad rates

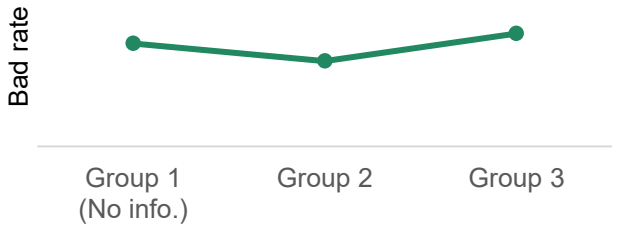
Time from the last time customers being in debt group 2



Expected trend: ⬆ Actual trend: ⬆

Customers with recent debt group 2 are worse

The highest debt group of customer



Expected trend: ⬆ Actual trend: ⬆

The higher debt group is, the worse customer is

5. Model performance

- Gini coefficient is used to assess model performance.
- The comparison of difference approaches on different products are showed in the table below.

Method	Credit card for User		UPL for User		Driver	
	Gini	PSI	Gini	PSI	Gini	PSI
Logistic regression	57.3%	1.7%	42.3%	3.9%	48.7%	5.9%
Gradient Boosting Tree	58.5%	5.0%	43.1%	7.8%	49.7%	8.1%
Recurrent Neural Network	59.7%	6.3%	44.1%	9.3%	50.7%	9.6%

➡ All the models have good predictive power.

Comment

- Ride-hailing data is effective in customer risk assessment.
- The logistic regression models are deployed in the system for decision making because
  - ✓ Logistic model is more transparent in term of model explanation
  - ✓ Logistic model is easy for implementation in Loan original system.
  - ✓ Other approaches don’t have significant higher performance and less stable than logistic regression.

6. Summary

1. Ride hailing is important data to predict credit worthiness of customers.
2. Ride hailing data is stronger than traditional data.
3. Ride hailing is correlated with Telco data (see Annex for comparison between Ride hailing data and Telco data).
4. Telco data is stronger than Ride hailing data.
5. It is needed to have more robust engine which allow to implement advanced algorithm to better predict credit worthiness of customer.

THANK YOU

Annex

Alternative data comparison

1. **The telco data is stronger** than ride hailing data for credit risk assessment. The main reason are:
   - ✓ Customers often have longer time in relationship with Telco provider than with online transportation platform.
   - ✓ Usage frequency of telco is higher than transportation.
   - ✓ Telco customers are more loyal than ride hailing customer.
   - ✓ Telco data is richer and higher diversity degree than ride hailing data because the requirement of telco data accuracy is higher than ride hailing data.

2. Below is the **characteristics of good customer** in view of **common features**. **All of them have the same trend in compare with bad rate**, except behavior in the weekend.

Key features	Telco data	Ride hailing data
Behavior at night	Number of outgoing SMS at night is small	Number of trip requests at night is small
Behavior on weekend	Fewer incoming call on weekend	Multiple drop-off location on weekend
Type of service, payment	No using fast credit (advance payment)	No payments by cash
Balance	High average main balance	High average main balance
Time of relationship with organization	Long time of relationship with organization	Long time of relationship with organization