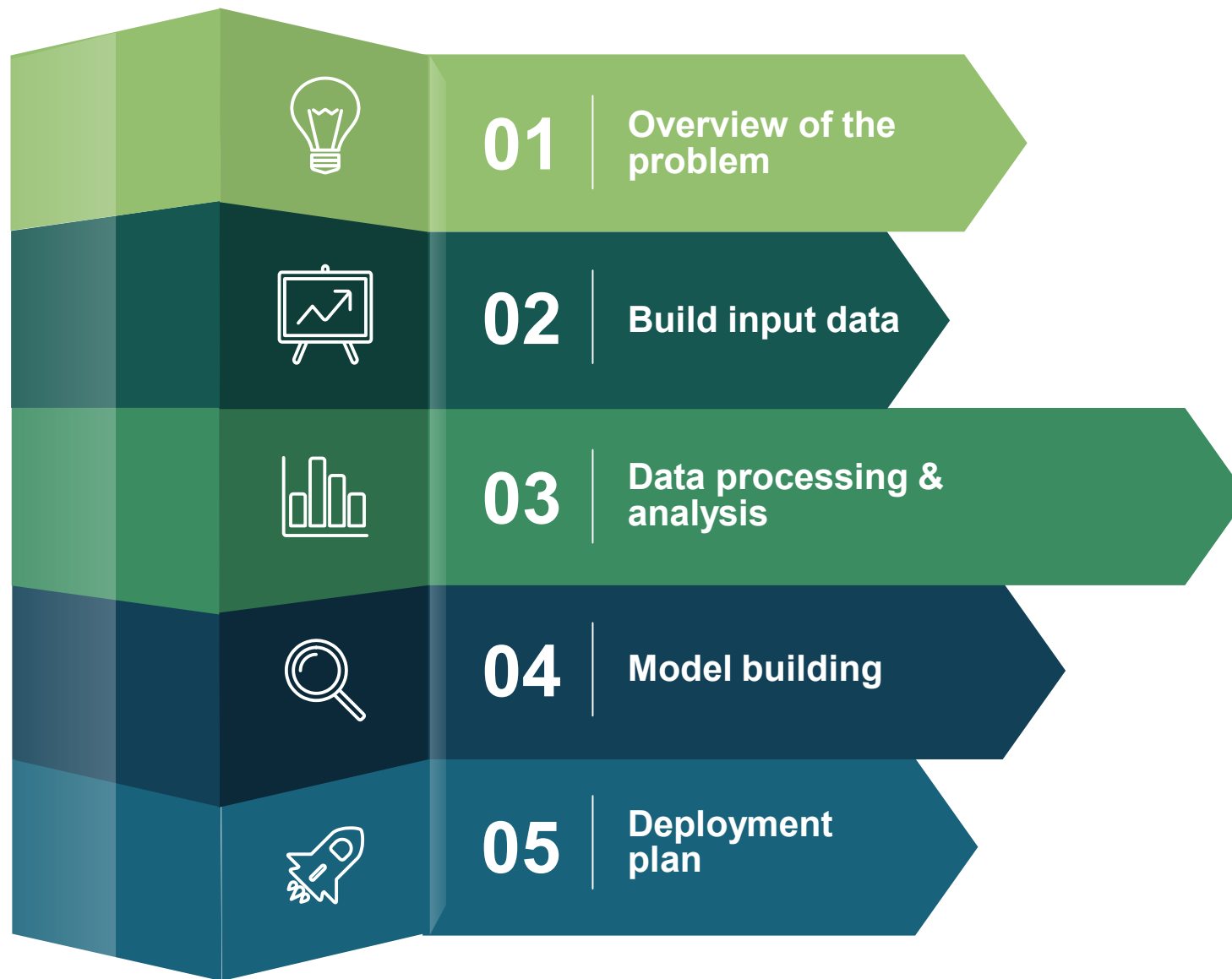




# CREDIT SCORE

A credit scoring model is a risk management tool that assesses the credit worthiness of a loan applicant by estimating her probability of default based on historical data. It uses numerical tools to rank order cases using data integrated into a single value that attempts to measure risk or credit worthiness.



1

# Overview of problem

Introduction of model

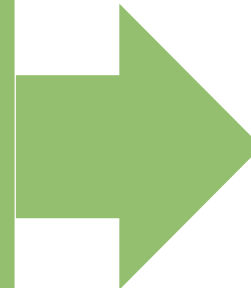


# Current model problem



80%

Borrowers with no credit history make it difficult to score credit using traditional methods\*



Many organizations have used telecommunications data to rank



trustingsocial



kt



Viettel, with the great advantage of holding 55% of the telecommunications market share in Vietnam, is able to help millions of Vietnamese people have credit scores, to access capital, and at the same time, help credit institutions reduce risks when Disbursement.

\* Theo dữ liệu báo cáo World Bank 2020



# Telecom data advantages



Wide coverage, all Viettel subscribers ~ 55% of Vietnam market share



The model-matching data is selected by the algorithm, reducing human subjectivity



Big data size with billions of records per day



Scoring algorithm is mostly machine learning, few assumptions and human intervention



Constantly updated data frequency with low latency

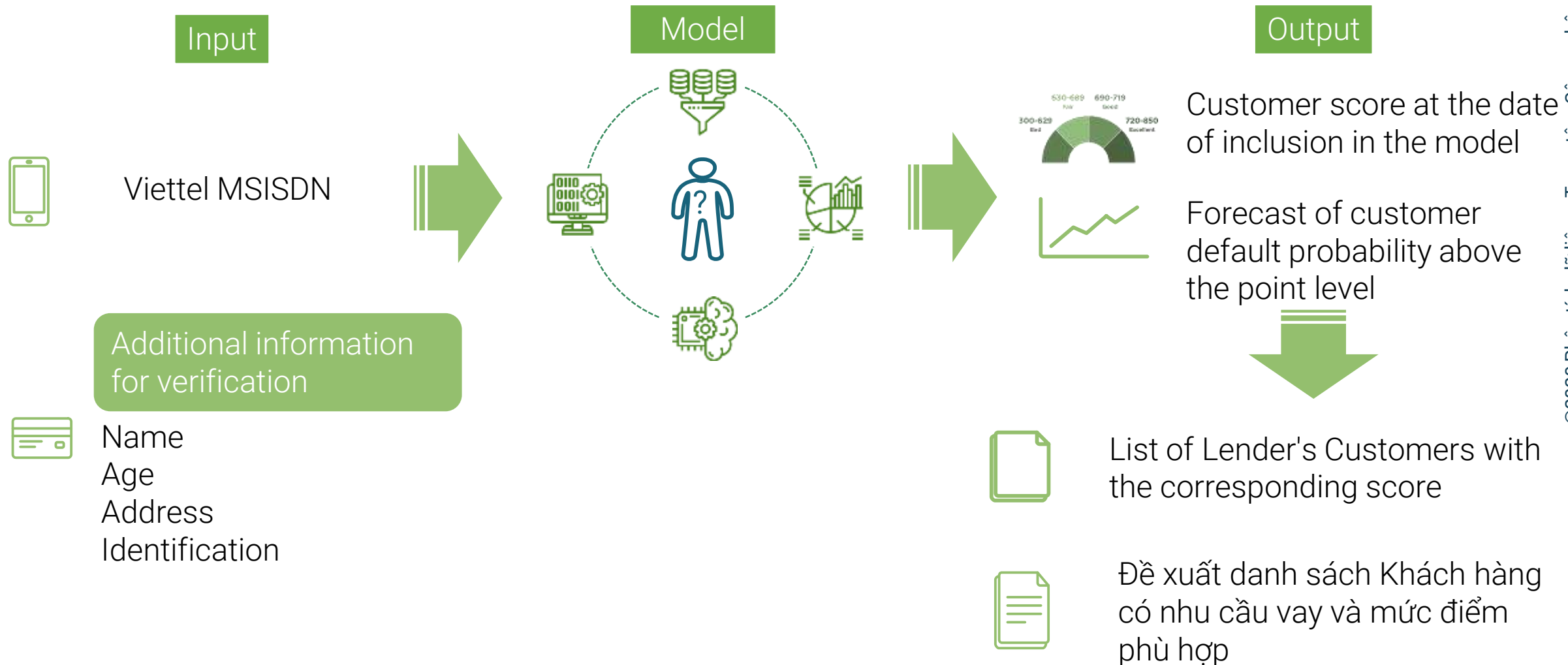


The quality of information is valuable for predicting credit scores, see pages 6 and 11 for details



# Overview of problem

Credit rating for customers with unsecured consumer loans. Illustrated model:



2

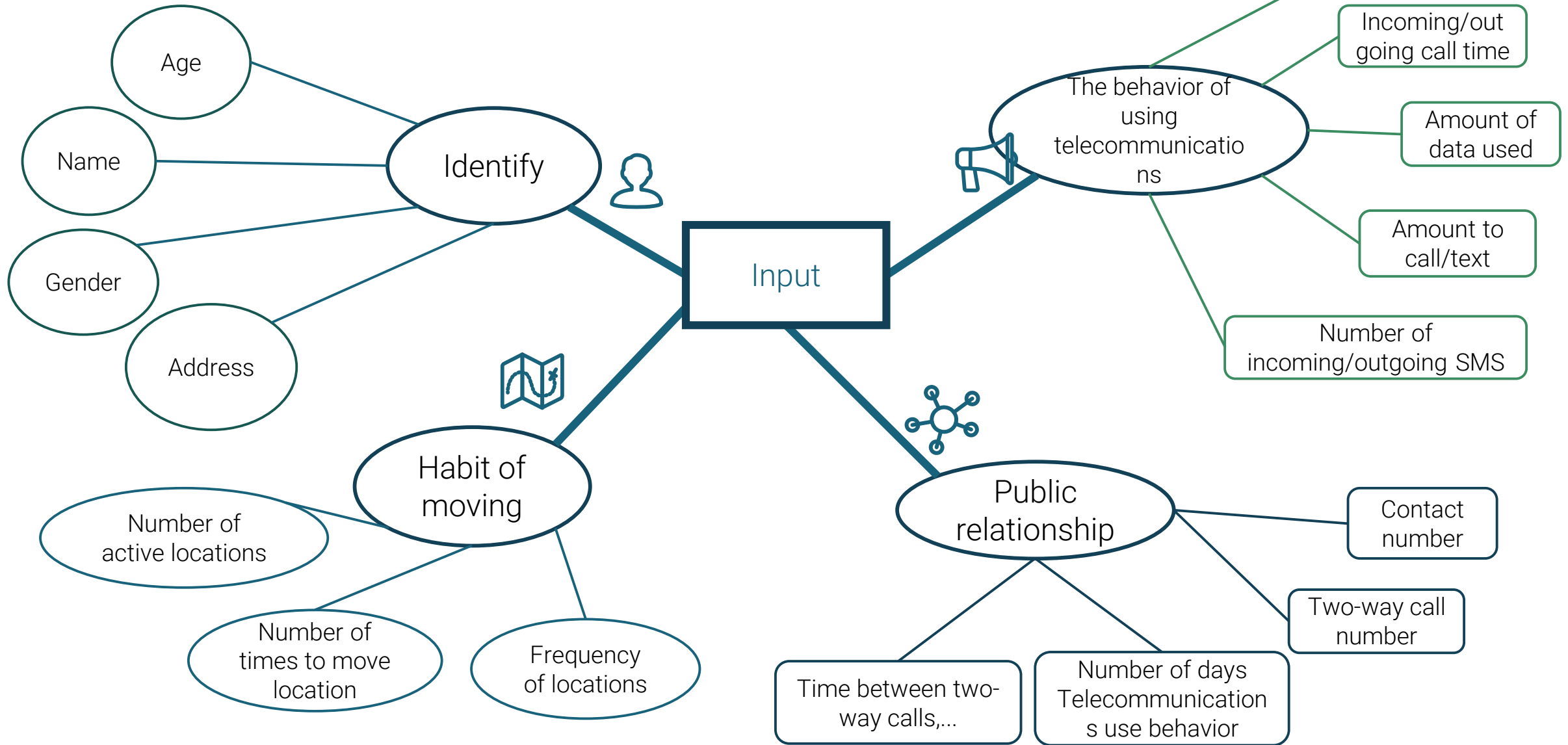
# Features Engineering

Features Engineering



## 2. Input data

# Input type







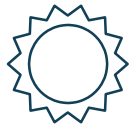
## 2. Lấy dữ liệu đầu vào

# Mô tả dữ liệu đầu vào

The total number of input variables is 1988 and the model data is updated weekly to ensure continuous model improvement. The problem uses 06 different time points:



### During the week (Monday - Friday)



Working (7h30 - 17h00)



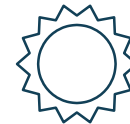
Rest (17:00 - 23:00)



Night (23h00 - 7h30)



### Weekend (Saturday - Sunday)



Working (7h30 - 17h00)



Rest (17:00 - 23:00)








Night (23h00 - 7h30)

- All variables containing telecommunications data are aggregated according to these 06 milestones to best distinguish customer behavioral characteristics.
- Data are extracted for 6 consecutive months: The increase in observation time is to enhance the stability of the observations and to see more clearly the trend of using telecommunications equipment of subscribers. In addition, this also contributes to partially preventing new junk sims from operating.



## 2. Input data

# Identity data

-  Age of customers: Age distribution is correlated with bad debt ratio (details on page 21)
-  Subscriber age: The older the subscriber, the more reliable it is, limiting the possibility of spam sims
-  Full name: Check the completeness of subscriber information
-  Subscriber's registered address: Regionality can also affect bad debt ratio (see page 22)
-  Gender: Usually, women have a lower bad debt ratio than men

➤ Subscriber with complete, unambiguous identity information will usually have more confidence



## 2. Input data

# Behavioral data of telecommunications use



### Data Mobile phone

Total amount of data used.  
Total amount of data uploaded  
Total amount of downloaded data  
Daily upload data volatility.  
Variation of download data by day



### ARPU

Total amount used  
Total amount for call  
Total amount for sms  
Total amount for data  
VAS  
Number of advances  
Total amount of advance  
Subscriber status: Before, after  
Number of times to change subscription status.



### Call/SMS

Total time of incoming/outgoing calls/sms, missed calls  
Variation of outgoing and incoming calls  
Frequency of calling/texting  
Number of missed calls  
Number of days subscribed to telecommunication activities

These data show the financial ability of the subscriber, the level of payment of the subscriber for telecommunications activities such as a full-paying and non-advanced phone subscription usually has a lower probability of default than the Other subscribers or postpaid subscribers usually have a better level of reliability than prepaid subscribers.

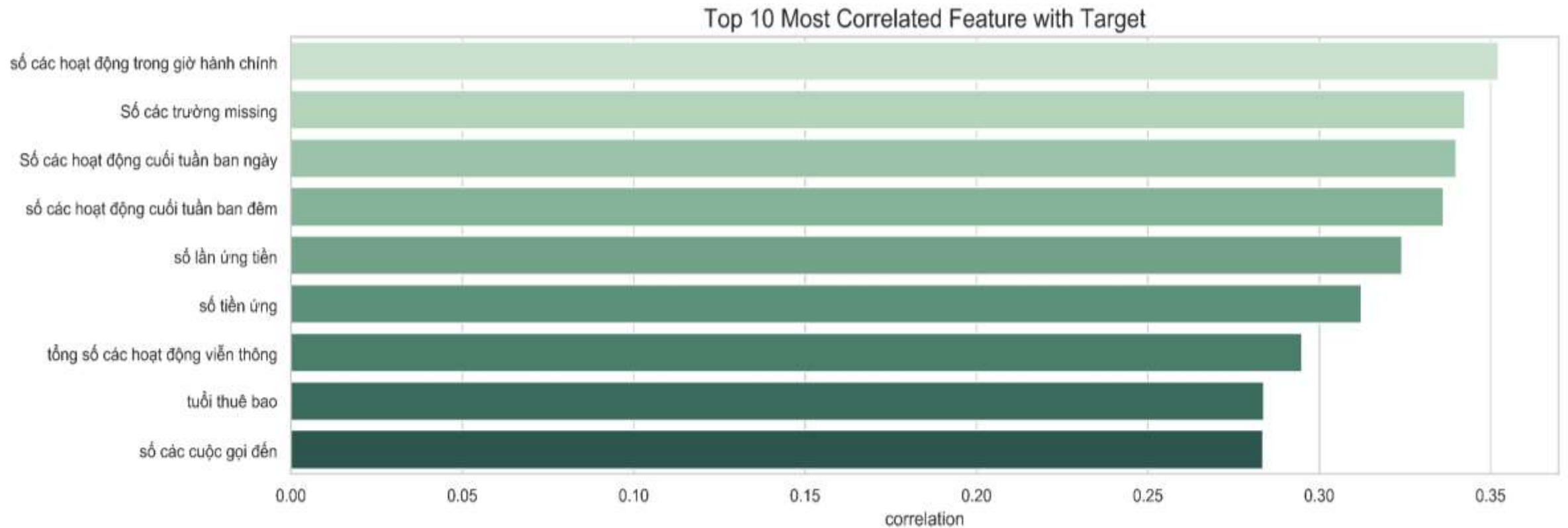
3

# Data processing & analysis

Data processing & analysis



# Correlation check



➤ The top 10 variables that are best correlated with bad/good debt labels have coefficients ranging from 0.28 to 0.35. It can be seen that these fields have a pretty good linear relationship with bad/good debt labels.



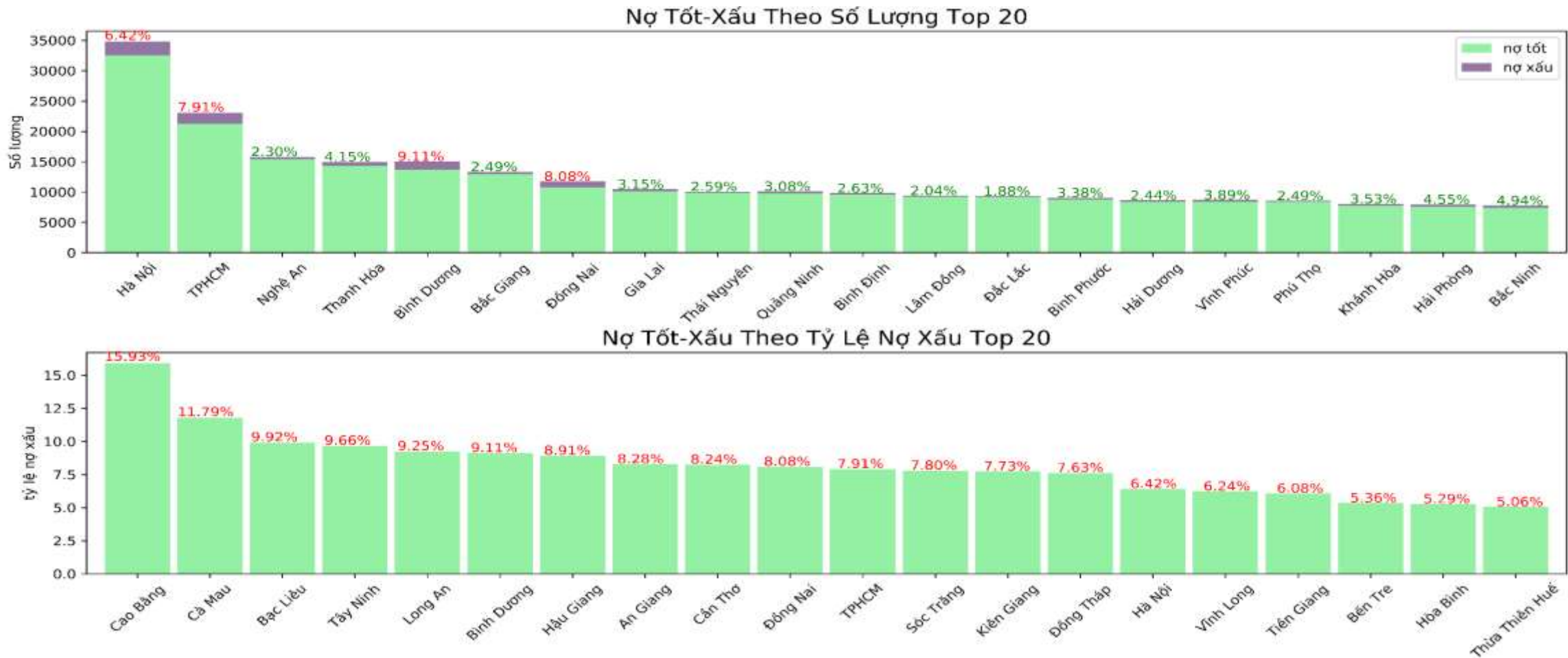
# Distribution of variable



➤ The bad debt ratio is high in the age group under 32 and much lower in the other group.



# Distribution of variable

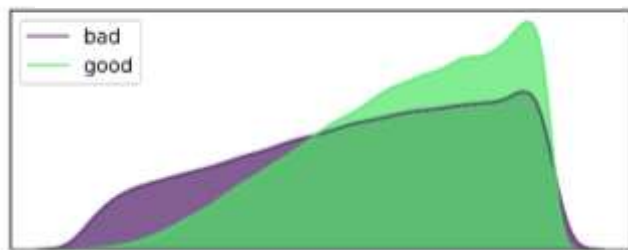


- Borrowing subscribers are distributed widely in big cities/provinces such as Hanoi, Ho Chi Minh City, Nghe An, Thanh Hoa, in which Hanoi, HCM, Binh Duong, Dong Nai have high bad debt ratio. mutation.
- The provinces of Cao Bang, Ca Mau, Bac Lieu,... ranked in the top of bad debt, but the number of observations is low, the reliability is not high.

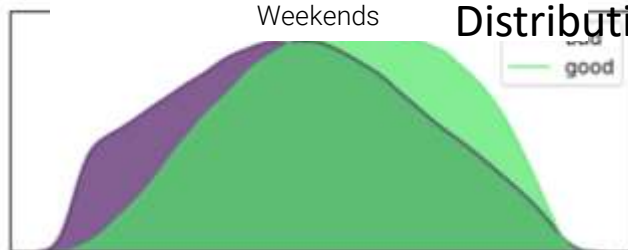


# Distribution of variable

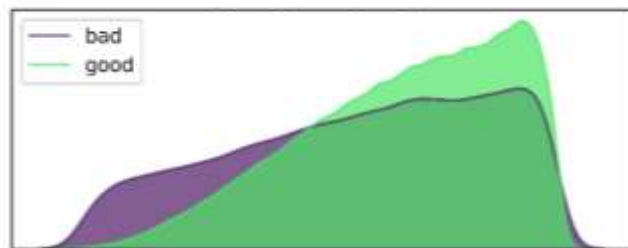
Average Distribution of Telecommunications Days in the Evening of the Week



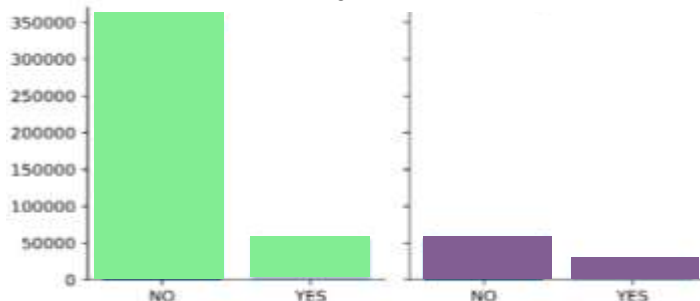
Average Distribution Of Frequency To The Most Popular Locations Evenings & Weekends



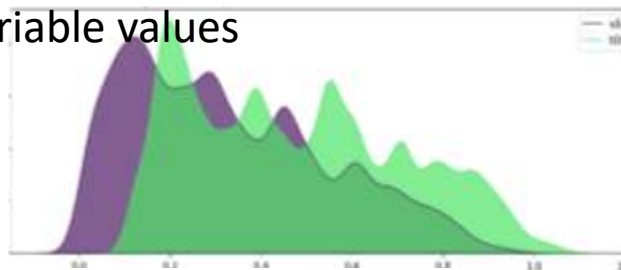
Average Distribution of Days Telecommunications Usage In The Evenings & Weekends



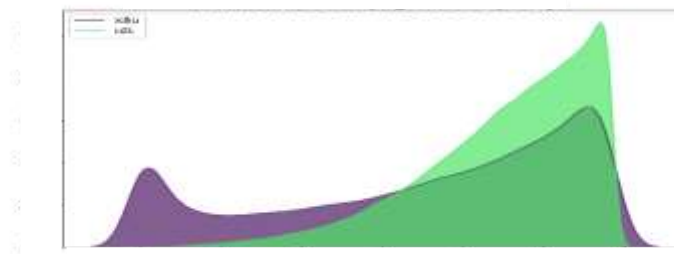
Distributing Subscribers At least 01 Time Change of Work Location



Telecommunications Subscriber Age Distribution (Varied Normalization)



Distribution of Number of Calls/Month of Subscribers



Distribution of variable values

Analysis of some important variables found that the distribution of values is discriminatory in terms of good and bad credit



Telecommunications data has the ability to give credit ratings to subscribers

Nợ xấu

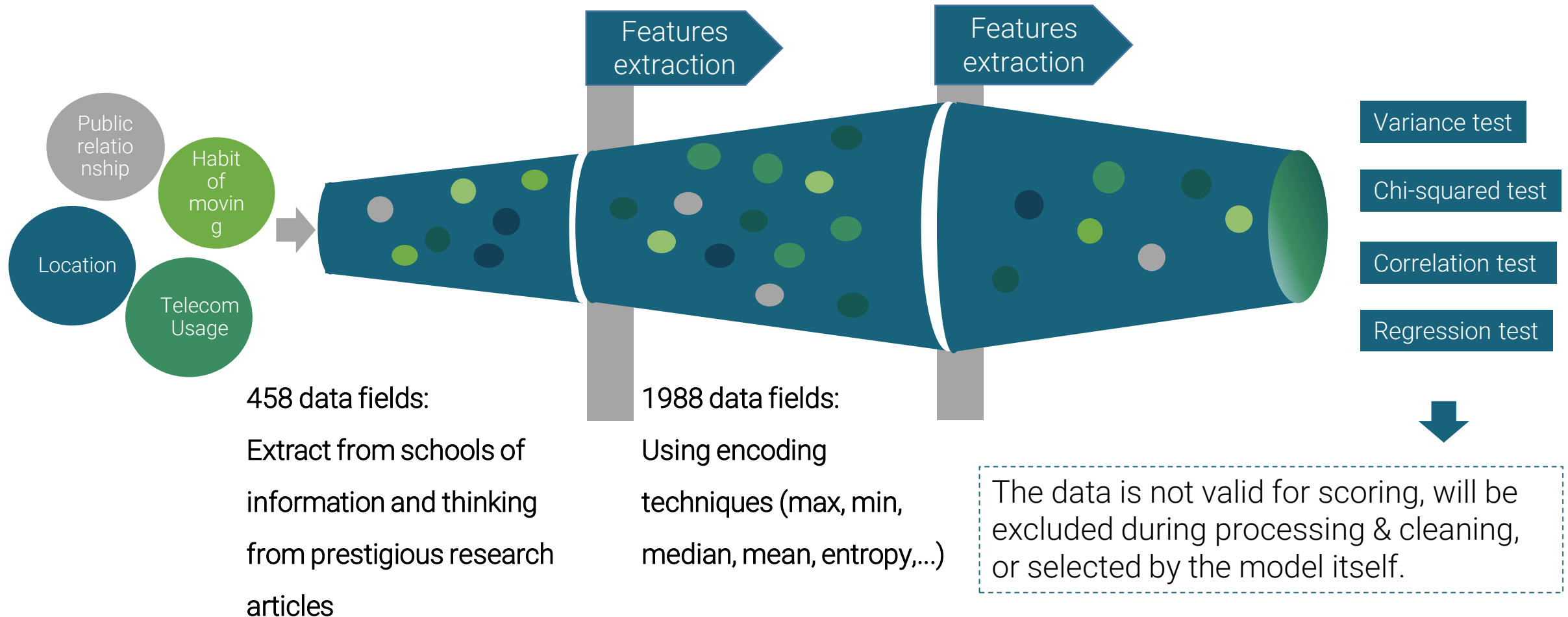
Nợ tốt





# Data selection mindset

Data generation thinking will be in the direction of maximizing exploitable information, including three steps: creating basic data (raw data), exploiting information fields from basic data (Features extraction), using transformation techniques for generating secondary data (feature extraction).



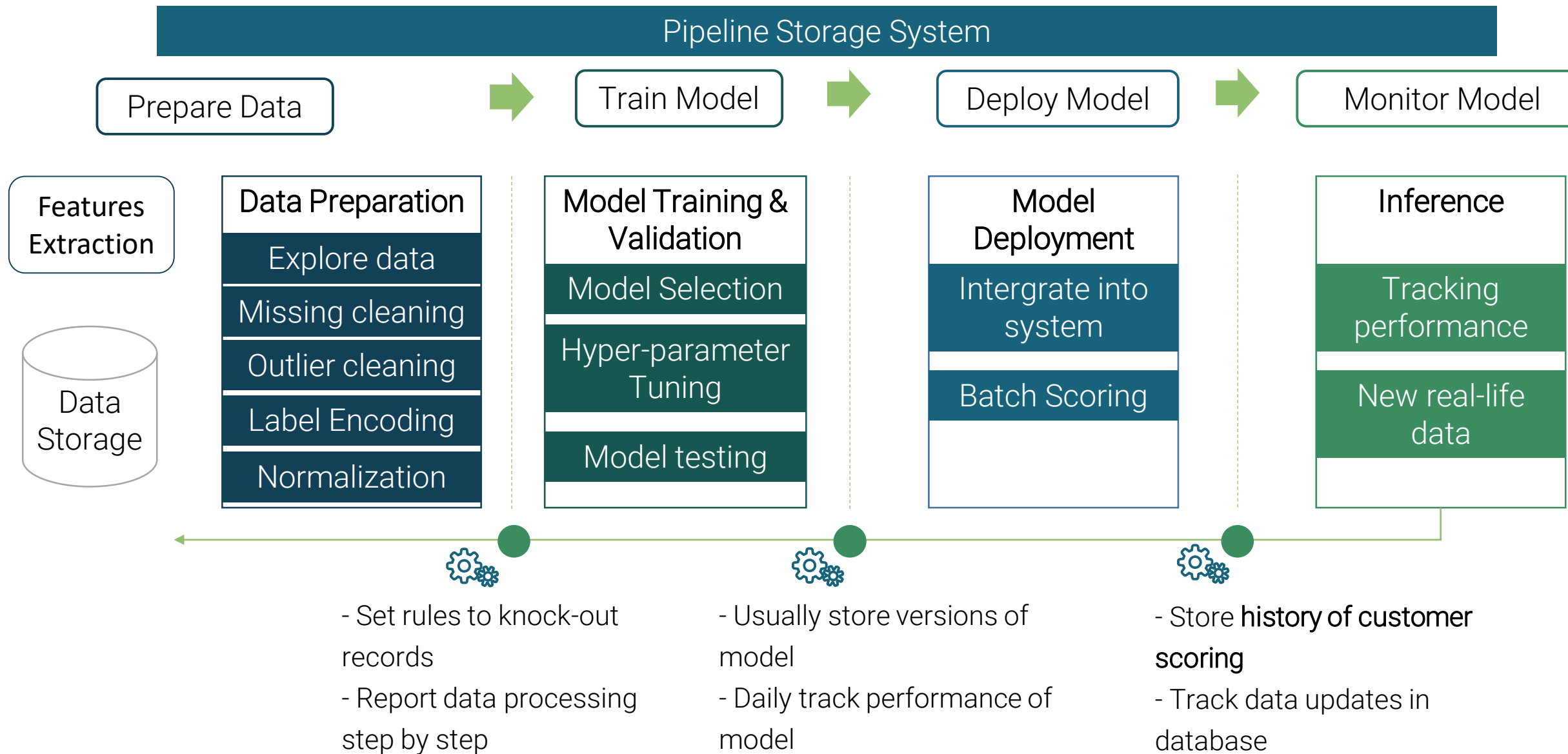
4

# Credit Score Model

Data modelling

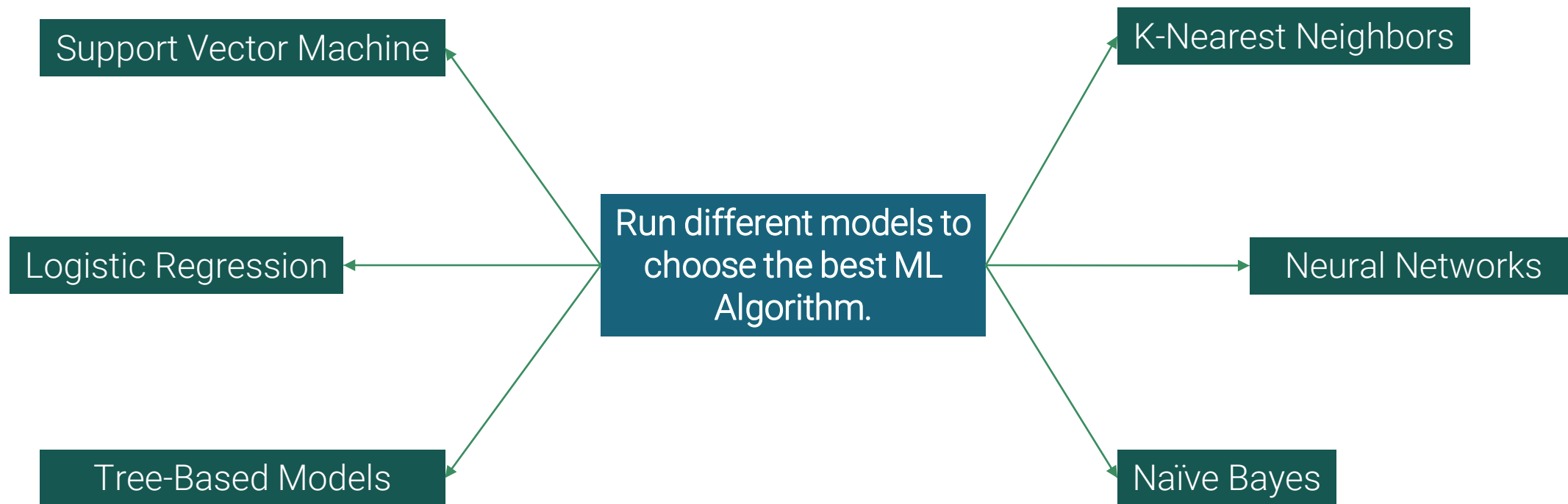


# Model flow overview





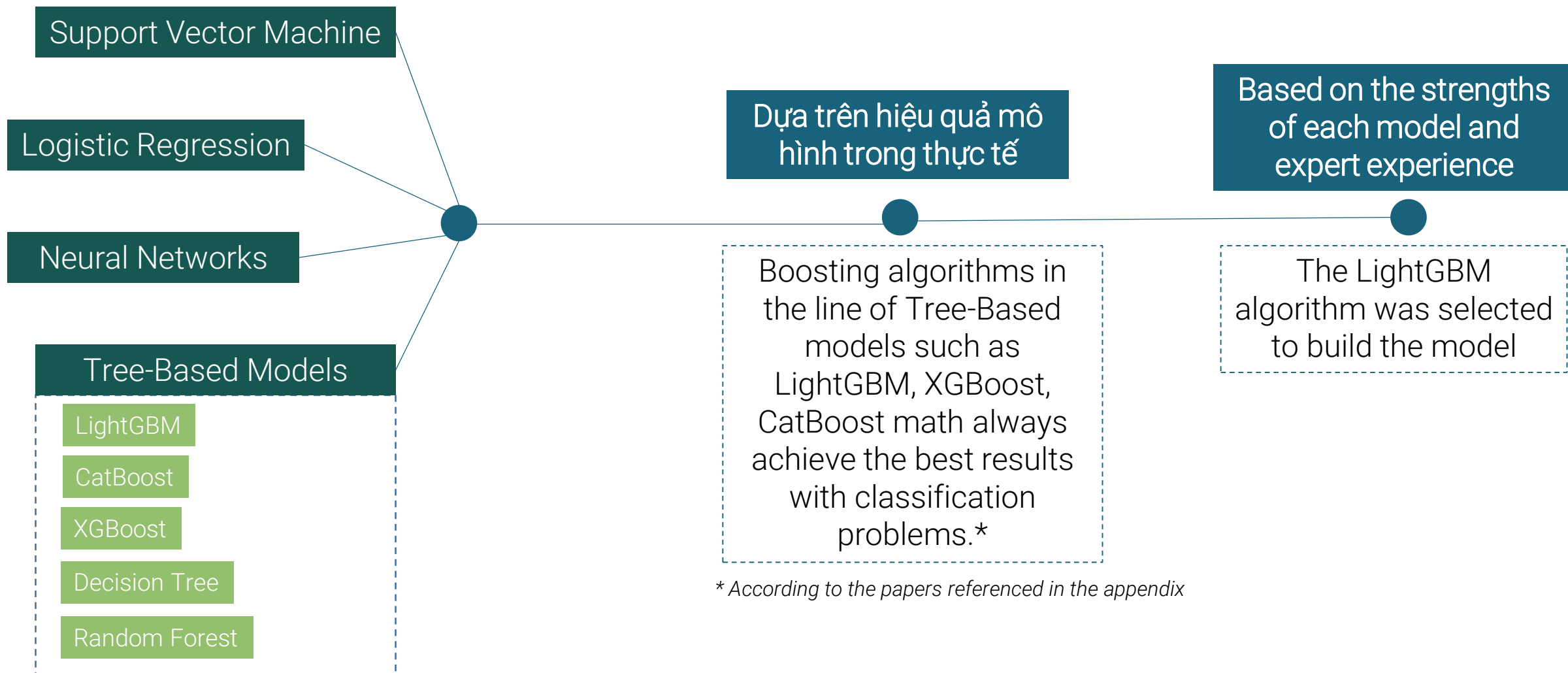
# ML Algorithm selection



*Simple algorithms like Logistic Regression or Naïve Bayes are used as baselines.  
For each algorithm, run bayesian research CV to find the parameter set that best fits that model*



# ML Algorithm selection



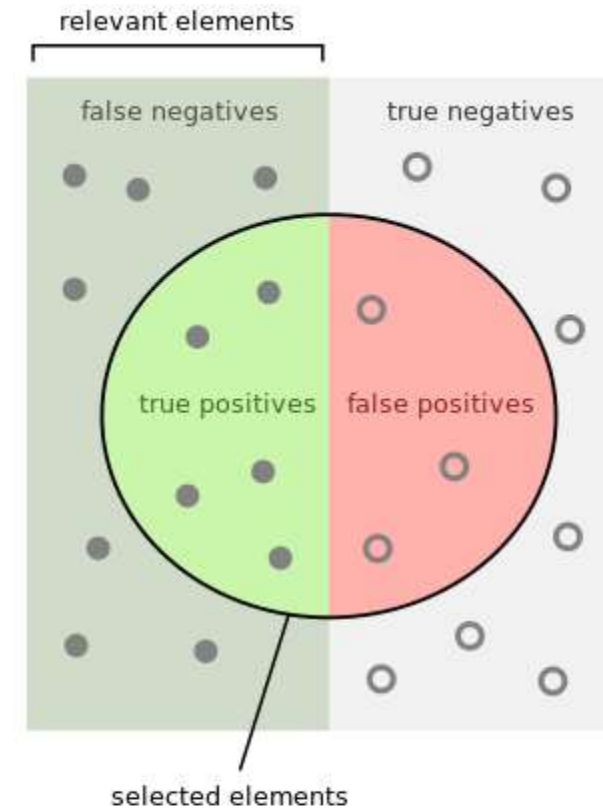
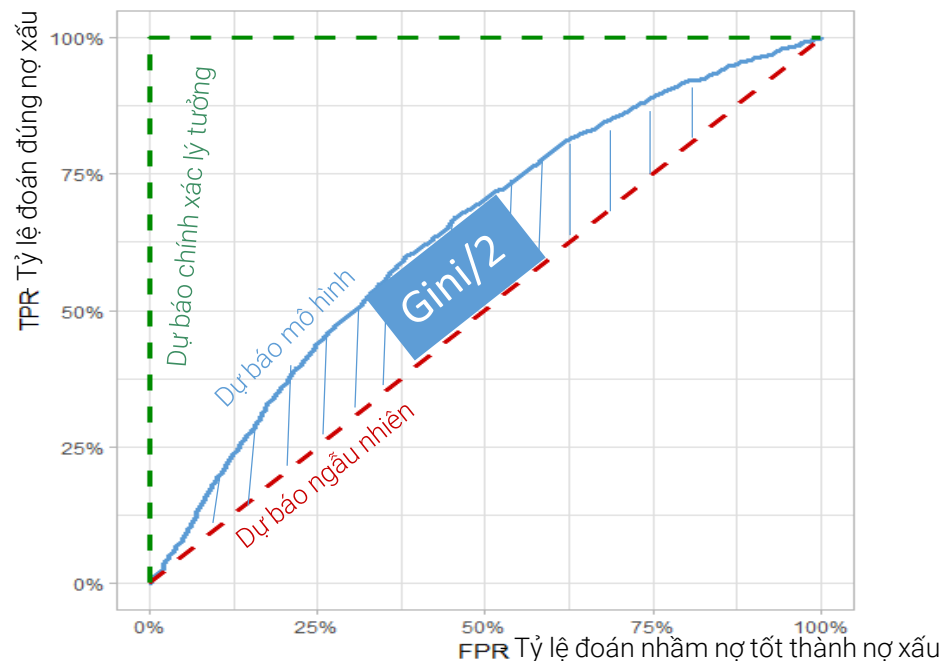
\* According to the papers referenced in the appendix

Basically, the algorithms Random Forest, LightGBM, XGBoost, CatBoost are the same, belong to the same family of boosting algorithms: Use many weak Tree Decisions to create a Robust Decision and limit overfitting.



# Criteria for measuring results

- Gini: This value ranges from -1 to 1, corresponding to completely wrong guess and absolute right.
- Recall: The ratio of correct prediction of bad labels to total number of bad labels.



How many selected items are relevant?

Precision =



How many relevant items are selected?

Recall =





# Prediction Results

	Bad MSISDN Prediction	Good MSIDN Prediction	Total
Actual bad MSISDN	16,483	4,327	20.810
Actual good MSISNG	211,030	443,801	654.831
Total	227.513	448.128	675.641

- Gini = 0.72 mean error 0.01.
- Balanced accuracy = 73.49 %  
average error 0.06%
- Recall score = 79.21% mean error  
0.04%

The structure of the confusion matrix can be changed by changing the model threshold for good/bad debt to better suit each situation:



⚠ Minimize bad debt: In case we don't have a lot of capital, or have a low risk appetite, it will lower the high threshold, but the limitation is that there are many good debts that are not considered for loans due to strict standards.

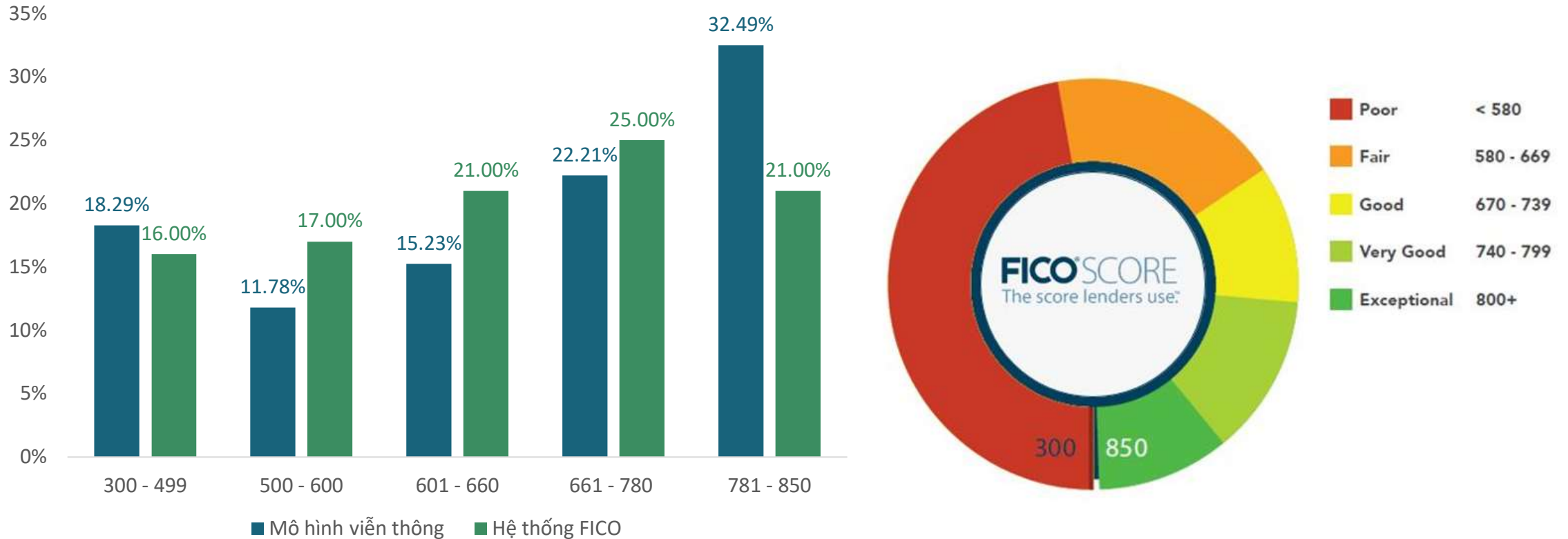


✓ Increase disbursement rate: In case we have abundant capital, and high risk appetite, consider a low threshold but the limitation is that the bad debt ratio is likely to increase.



# Compare model score spectrum

The forecast model provides a threshold of customer credit rating according to FICO standard (300 – 850) with 05 levels:



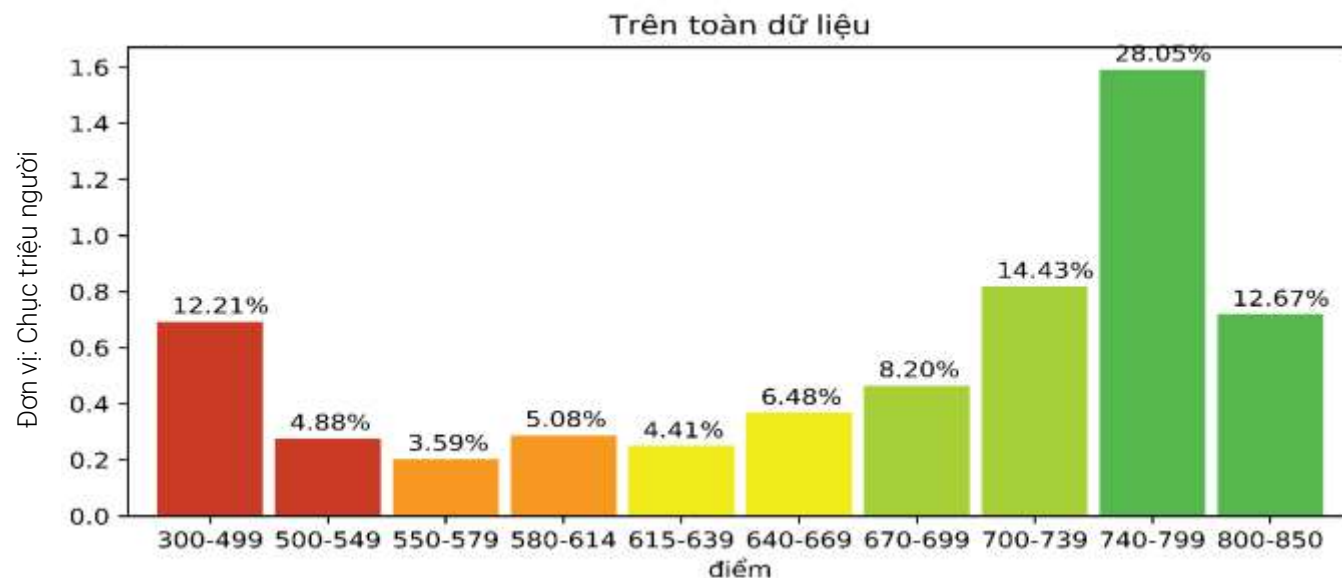
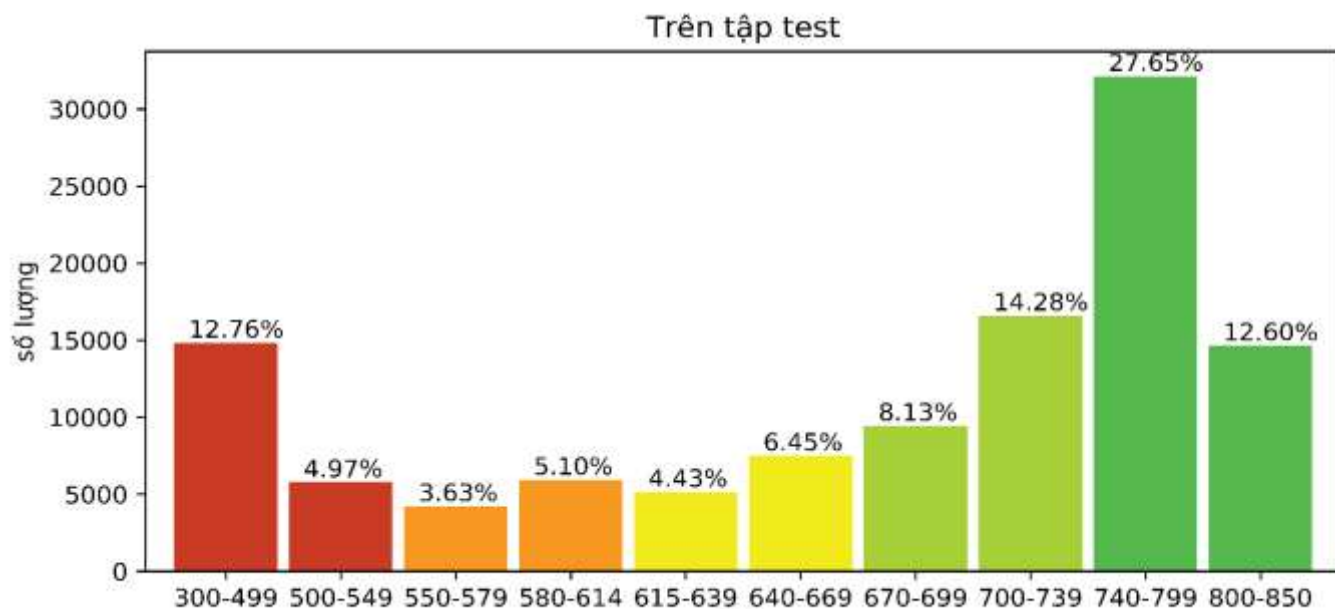
➤ The spectral distribution predicted by the telecommunications model has many similarities with the point spectrum distribution of the FICO\* system, showing the reliability of the model.

\* [Truy cập đường link xem thêm về tính phổ biến của hệ thống FICO](#)





# Compare model score spectrum

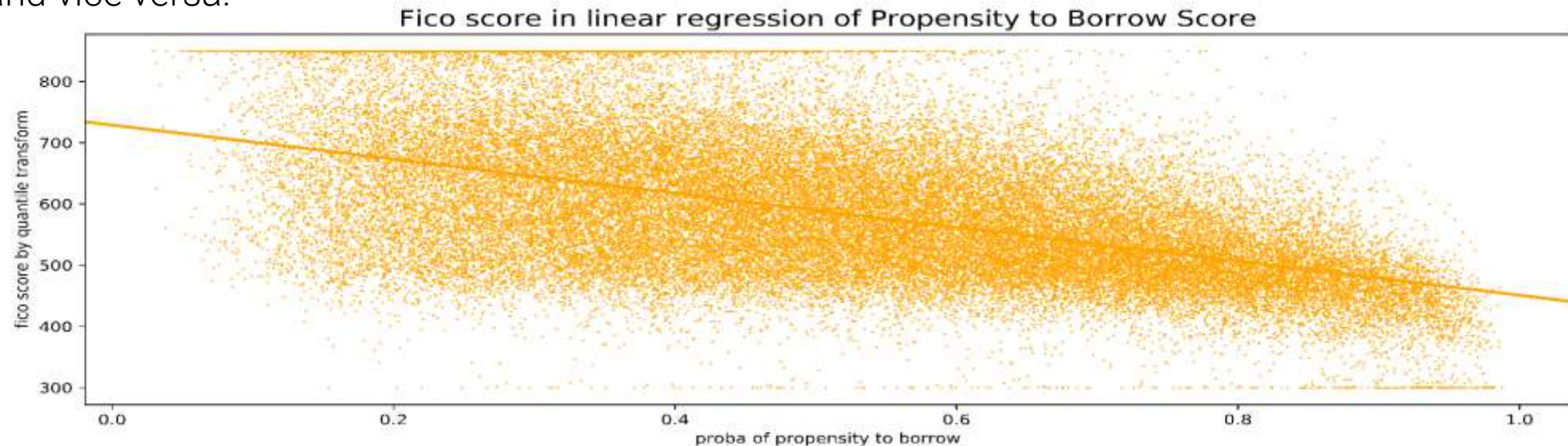


The forecast model provides the distribution of model scores on the entire data set of Viettel's telecommunications customers



# Result of loan demand score

Preliminary results of the model to forecast demand for borrowing (propensity to borrow). We see that these two points are inversely proportional, proving that subscribers with high Fico scores often have low borrowing needs and vice versa.



- Past Application: Running lead sets for EasyVay, SaveNow, Vietlott products
- The next step could be using a multi-product support model for marketing.

# Reference studies

- Lane, M., Carpenter, L., Whitted, T., Blinn, J.: Scan line methods for displaying parametrically defined surfaces. *Communications ACM* 23(1) (1980)
- Ding, C.H.Q., Peng, H.: Minimum redundancy dữ liệu selection from microarray gene expression data. *J. Bioinformatics and Comp. Biol.* 3(2) (2005) 185–206
- Prediction of Socioeconomic Levels Using Cell Phone Records
- Credit Scoring for Good Enhancing Financial Inclusion with Smart(2019)
- Business Intelligence Applications and Data Mining Methods in Telecommunications: A Literature Review [https://core.ac.uk/tải\\_xuống/pdf/71542208.pdf](https://core.ac.uk/tải_xuống/pdf/71542208.pdf)
- And many other at: <http://confluence.digital.vn/x/ZyRDAg>

# Notes

- Using data for 5 months instead of 3 months can add up to about 0.03 gini index. The author proposes to use data for 6 consecutive months to increase stability.
- Standard dropout from the model: Subscribers have at least 1 month without incurring telecommunications transactions. When running in reality, it is possible to issue a minimum score for subscribers who have interrupted telecommunication transactions (incurred in month  $t$  but not in month  $t+1$ )
- With numeric data: Group into groups according to different months, turn into mean and std data of these data. These actions increased the gini index by about 0.015 compared to using the original data.
- For data categoricals: Encoder with label encoder normally. In the author's experience, encoder by other methods such as onehotencoder, target encoder does not improve accuracy if boosting algorithm is used.



# ML Algorithms Advantages

## Advantages of LightGBM

Low resource consumption, easy to run on Hadoop Spark (server) or personal machine (local).

Good results on many different datasets with very few parameter settings.

Good limitation of overfitting problem.

Easily deal with missing data, non-numeric data, and mislabelling.



## Advantages of K-folds\*

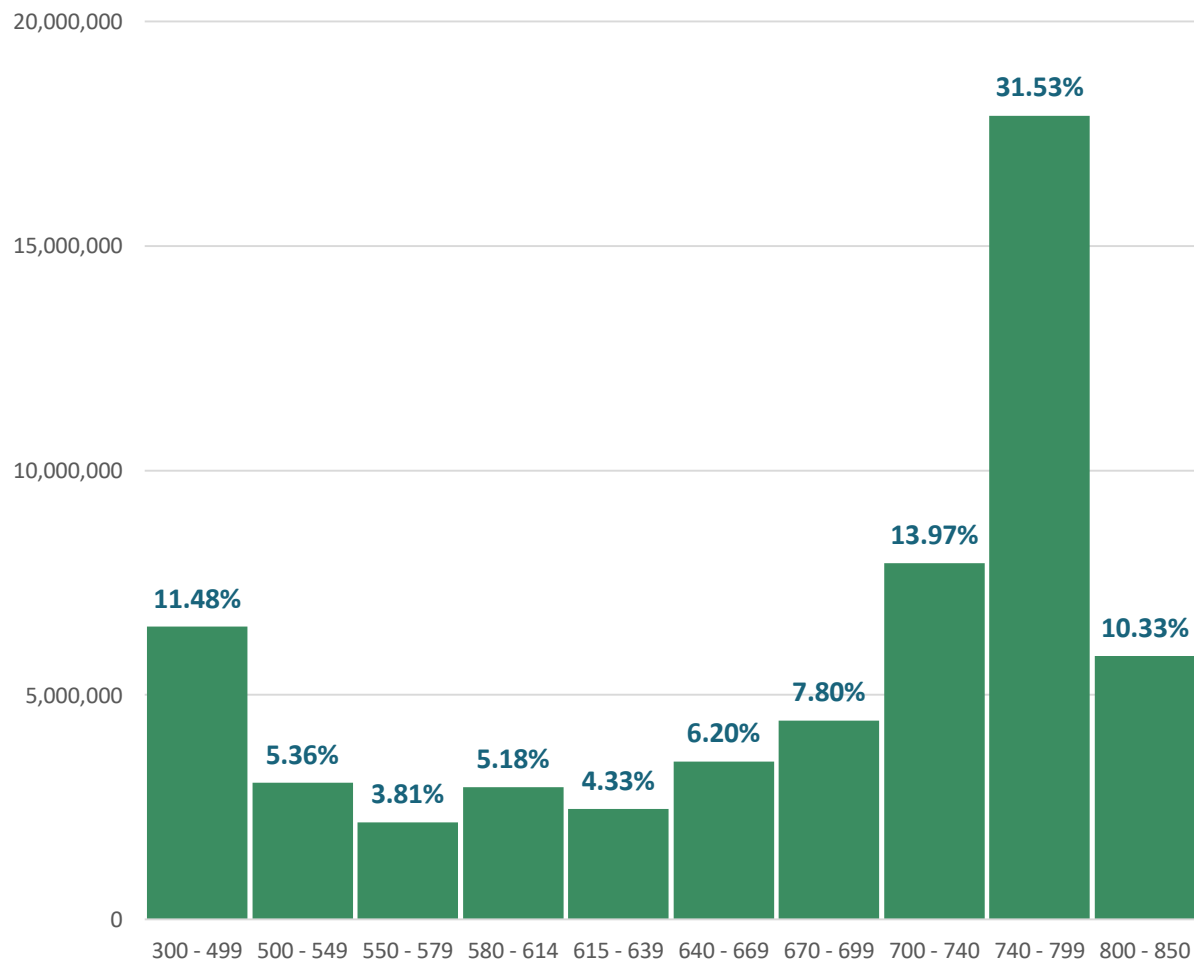
Applying K-Folds next to LightGBM aims to give the most objective and stable results

Minimize the confounding factors caused by the process of cutting the training and test sets.

Method: Randomly divide the data set into 5 equal parts. Train on 4 parts in turn and test the results on the remaining set

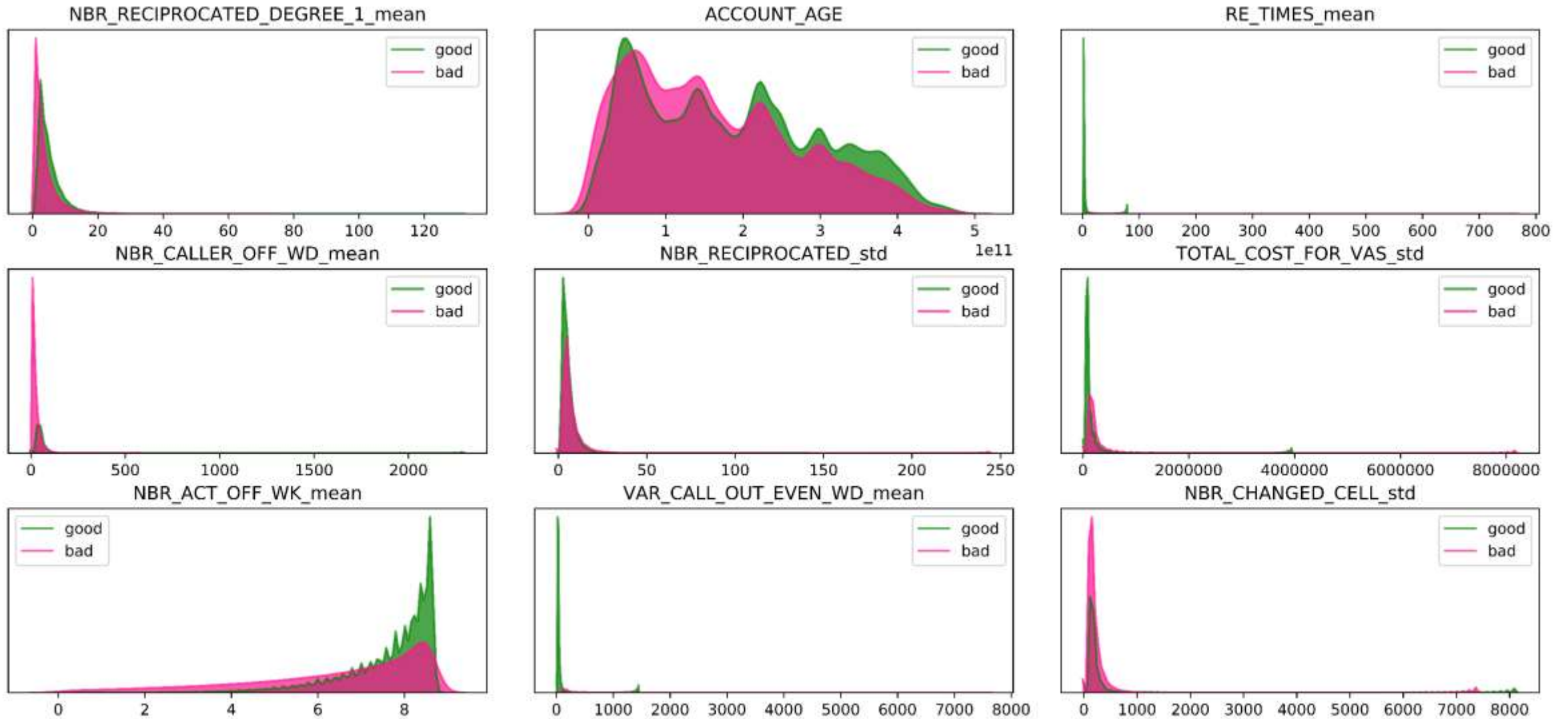


# Model point distribution



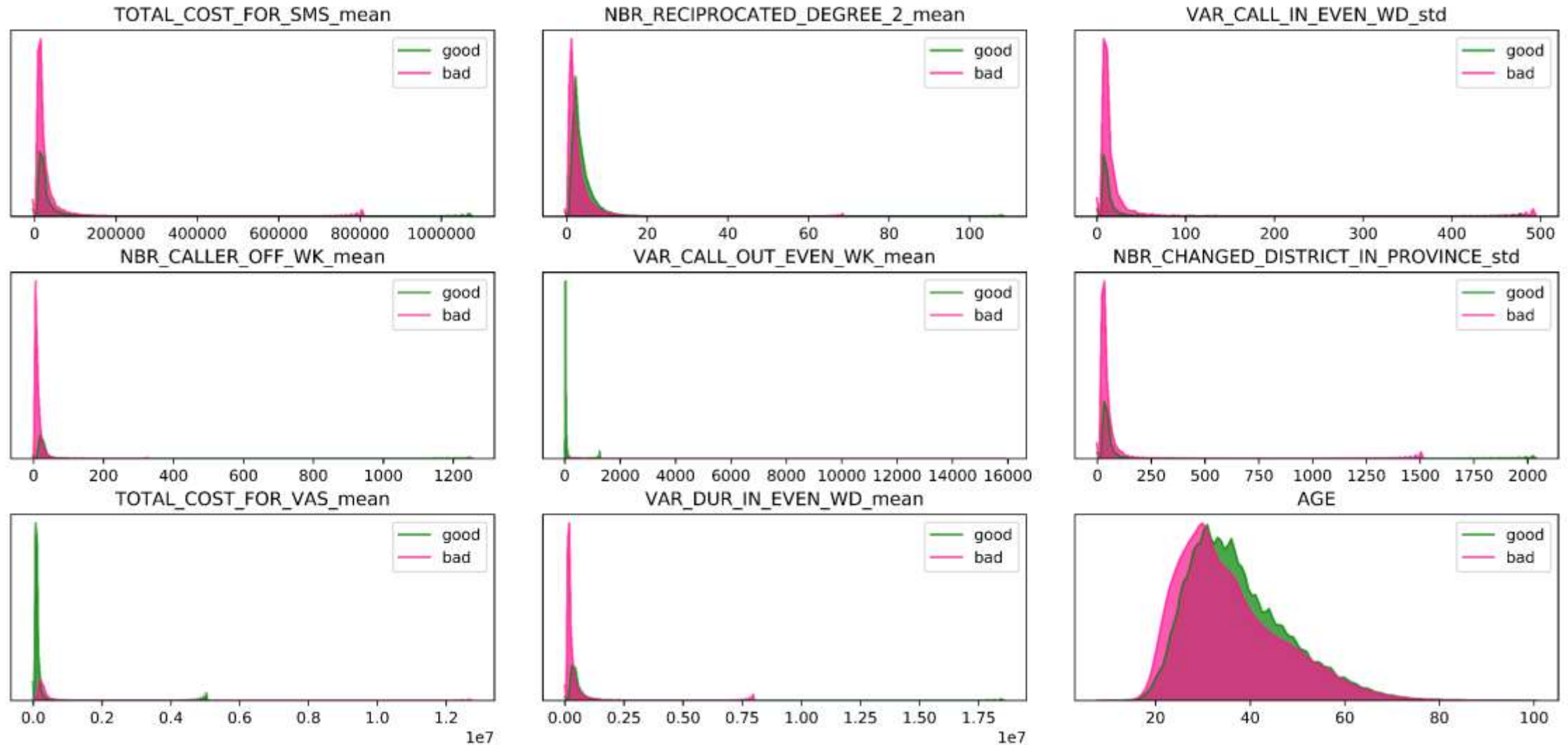
The forecast model provides the distribution of model scores on the entire data set of Viettel's telecommunications customers

# Distribution of variables



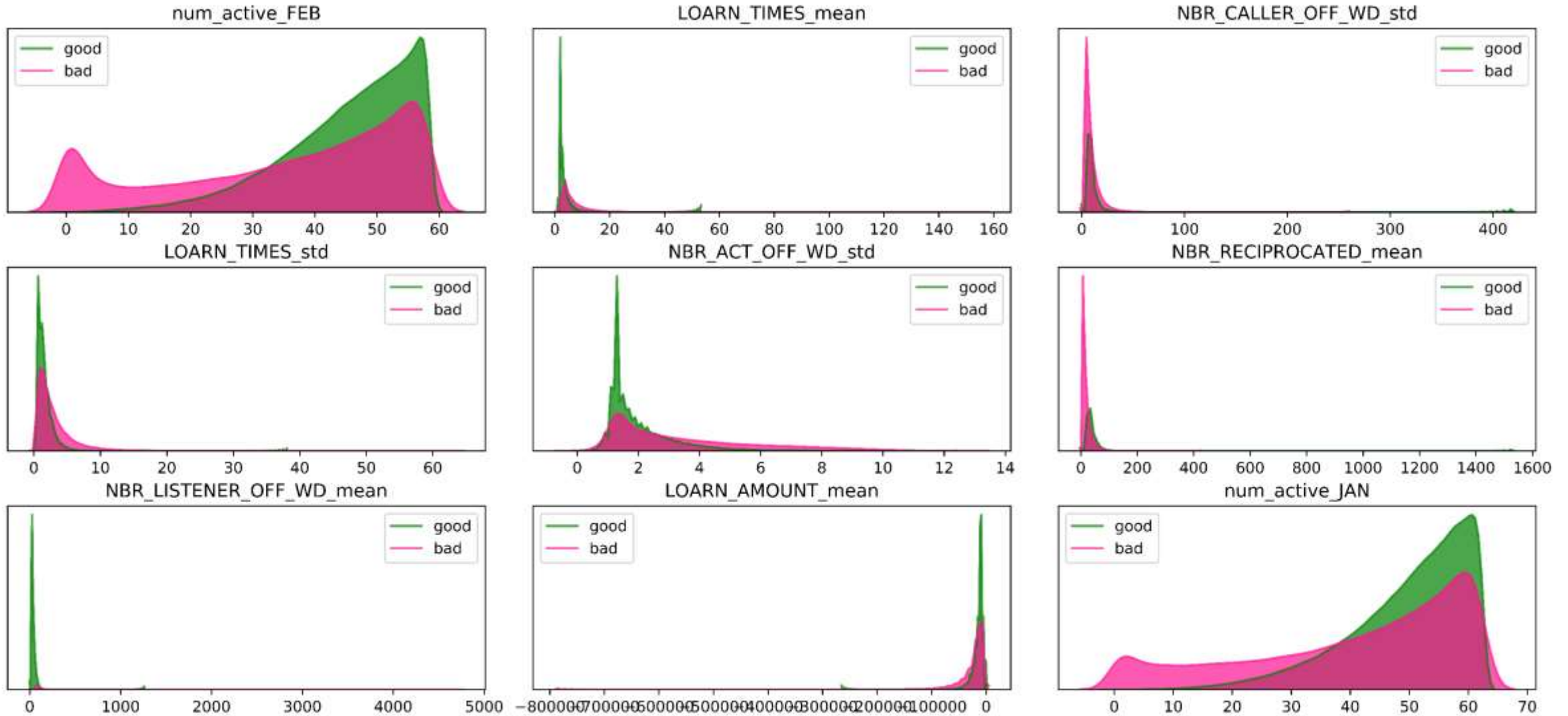


# Distribution of variables

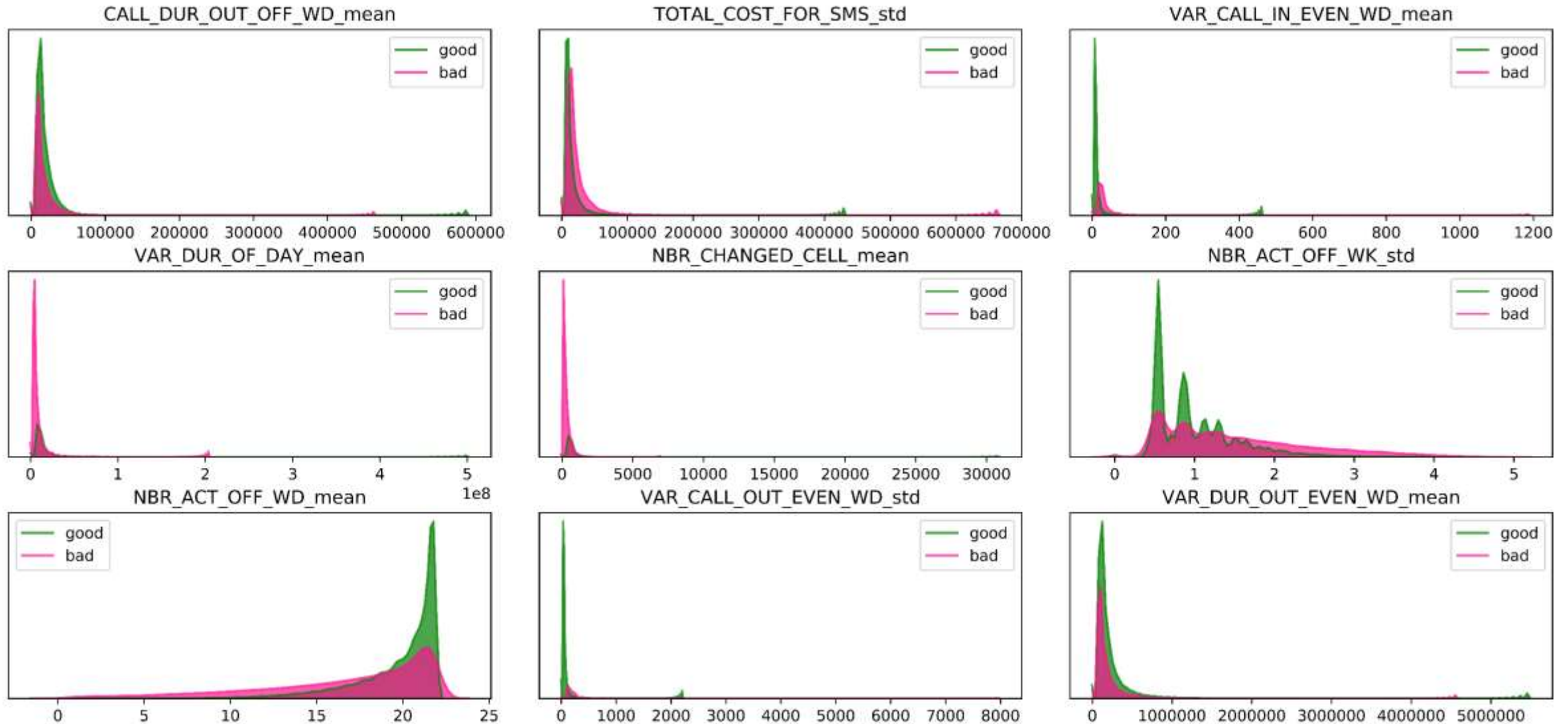




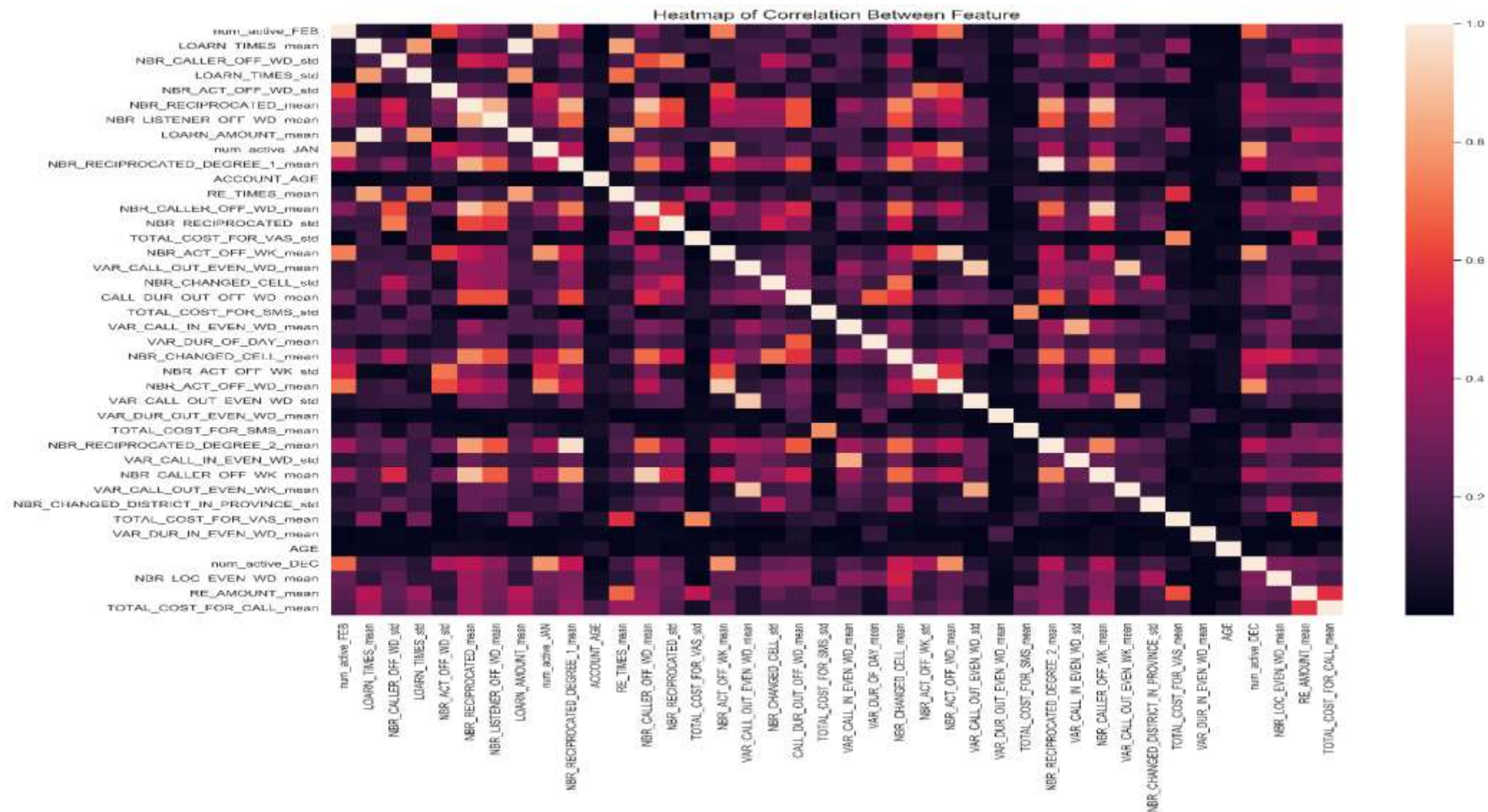
# Distribution of variables



# Distribution of variables



# Correlation coefficient heatmap



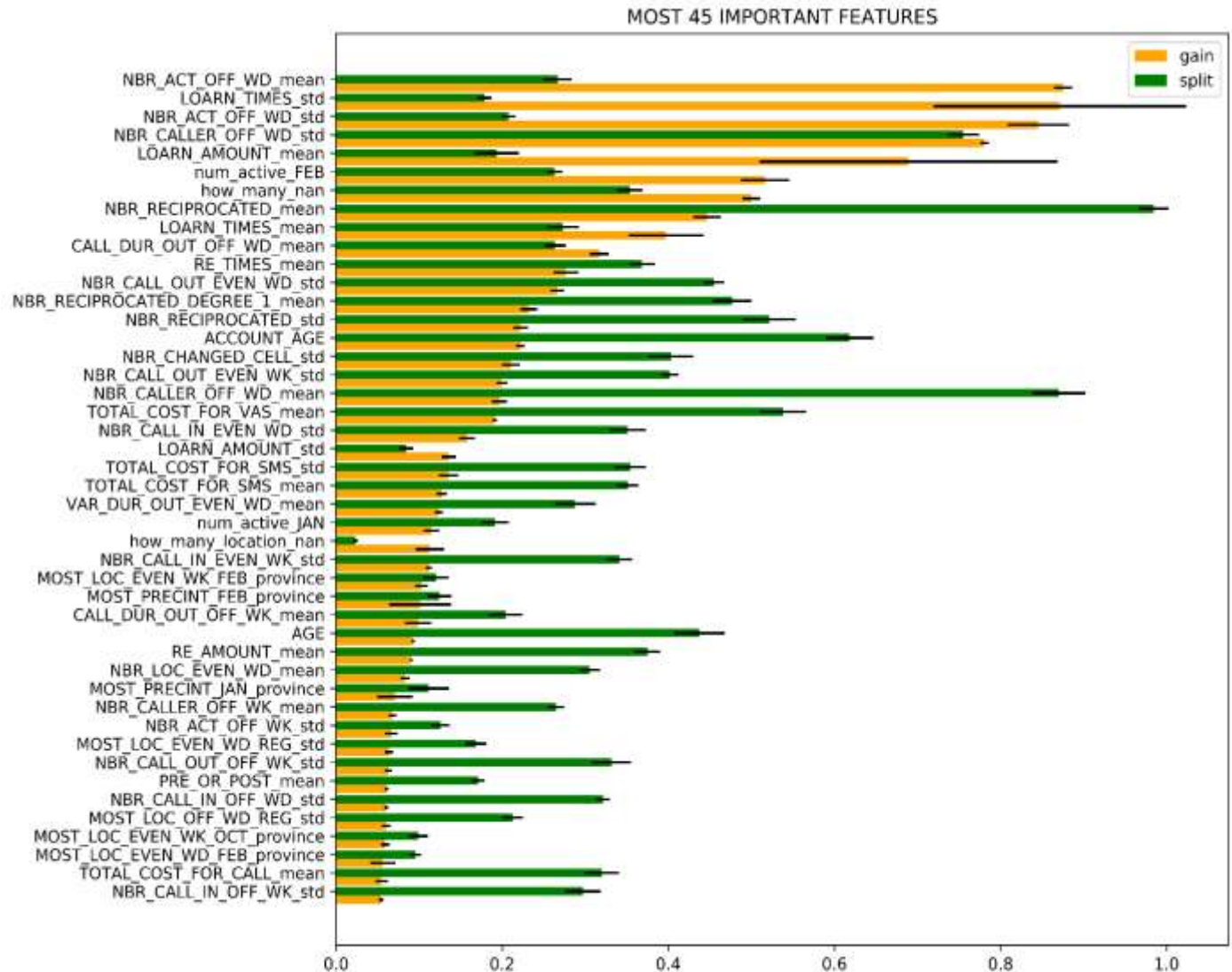
# Result evaluation

The model's plausibility is partially reflected in the Top of the Most Important Lists. Here we have:

Group data related to call frequency during office hours.

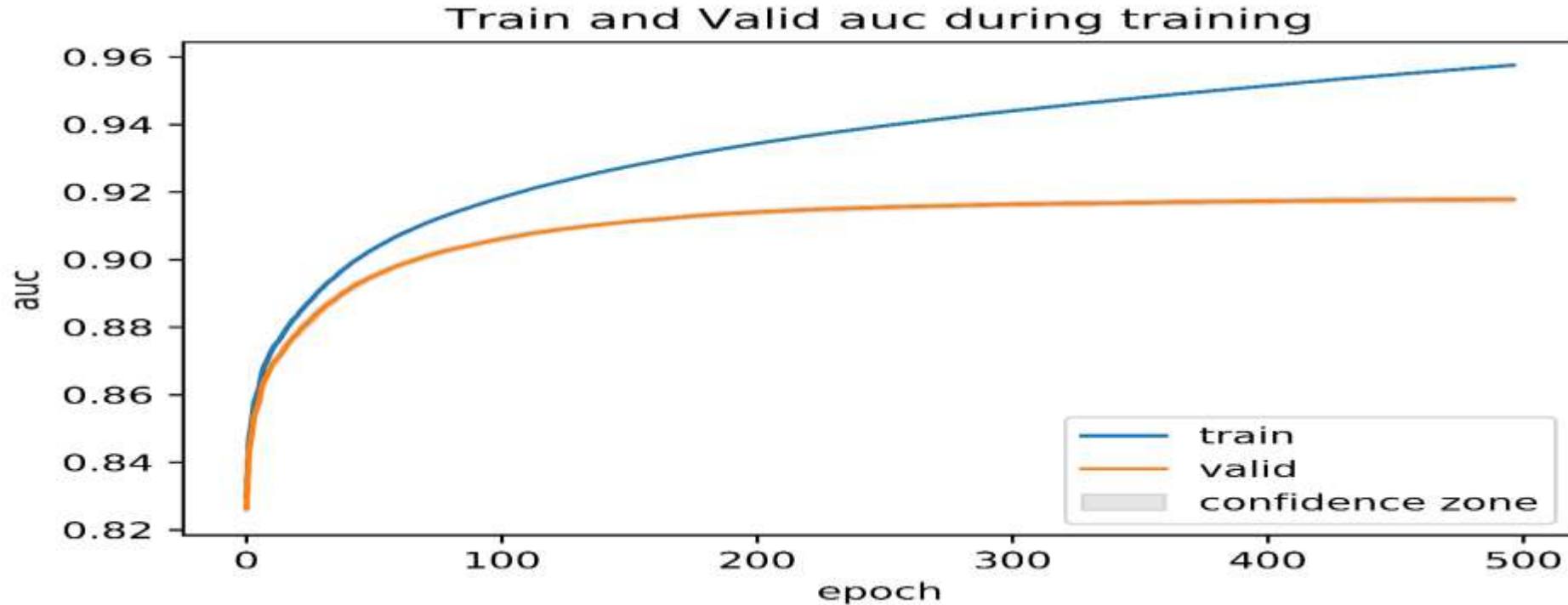
Group of data related to the frequency of telecommunications advances.

Groups are concerned with the subscriber's two-way interactions.





# Result evaluation



The models performed on different train and test sets all give very small standard deviation results, proving the very high stability of these models, because the dataset is large enough.