The income prediction module of the retail store's network

Kristina Pakhomova[1], Pavel Peresunko[2], Sergey Videnin[3] and Eugenia Soroka[4]
Institute of Space and Information Technology, Sibirian Federal University. Russia
e-mail:
[1]kpakhomova@sfu-kras.ru,[2]peres94@yandex.ru,[3]svidenin@sfu-kras.ru,[4]jollot@yandex.ru

Abstract

The main idea of this paper focused on the development of a program module, which predicts the pharmacy retail income by the machine learning theory. Beyond that, we want to introduce the best prediction model, which has learned by specific retail dataset. Notice, the architecture of program involves dynamic upload dataset, by Yandex" Internet service. The dataset represents the set of features and set of retail points, however in this task, the features describe the pharmacy industry. So on the first step will analyze the dataset and found out the correlation of the features. Next, will select the relevant features, which affect on income rate of the retail point. The last one will introduce to the prediction income Average model. In the last, will compare the three models, there is Average model, Gradient Boosting Regression and Random Forest Regression.

Keywords: Feature selection, Income prediction, Machine Learning, Artificial Intelligence

1 Introduction

Nowadays, Artificial Intelligence tools have become more actual in a business area, particularly machine learning approaches provide insightful economic analysis and an increase of company income. Indeed, business needs on the statistics, economics, marketing, and mathematical approach, for the analysis of the important retail features and profit markup. Sometimes the businessmen have not sufficient information about the features quality, which affects the profitability of the retail stores. Traditionally, the efficiency of features was calculated or analyzed by experience way. However, in the context of selecting the new retail store with an estimated future income rate, the businessman may make mistake. The experts need to aggregate a stores information (humans traffic, income rate, count of bus or metro station, business competition and etc) for drilling down a database of features. So, identifying the important and relevant features has a number of an advantage today.

Actually, machine learning provides a variety of choice of universal algorithms focused on feature selection, recommendation, prediction, and other tasks. Anyway, those algorithms do not provide the reliable result in a case a small sampling of fewer than 100 units and features count more than 120 units. Previous, the task of prediction income of retail was investigated by [5, 3]. By the way, the most advanced method is neural network [1], where authors introduce the simple linear regression then multiple regression and after describe the algorithm of neural network, which applied on a retail dataset from "Google Places API". In this paper we want to introduce our decision for prediction retail store income and compare it with the general machine learning methods.

This paper introduces the module for the retail store's income prediction based on the mathematical model with the specific dataset. In section 2, we will describe the prediction algorithms and introduce the Average method. In section 3, we will describe the dataset, particularly the set of features and their multiplicity. By the way, the huge number of features according to their relevant selection is scaling each feature to a given range. In the last section, we will apply the method to the dataset and compare the Average method with other general prediction algorithms.

2 The module development

Every store in the retail network has a set of features which influence on the successfulness of the store. By the way, for each retail store network has specific particular qualities describe an income of store. In this paper we want to introduce the module, which analyses the features of the stores' network, select only relevant features and predict the income for a future stores for target retail network.

For this reason, the Figure 1 shows the general mechanism to this paper, where we will introduce the prediction module, which runs a red contour. For the experiment the dataset was obtained by "Yandex" services and custumer's statistical informational.

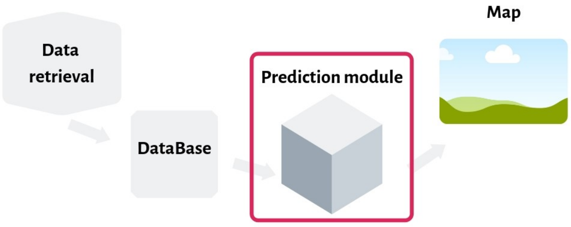


Figure 1: The diagram of the problem conception

The Prediction module includes the main three step, where on the first we normalize the dataset, then selected the relevant features by Boruta alghoritm and in the end to choose the best prediction alghorithm for set a task.

The raw data should be uniform in the case of the heterogeneous dataset. Indeed, exist a several type of features, where a dataset may have a binary type, counter and value. For this reason, we apply the min-max scaling procedure in the interval [0,1].

On the second step, we want to reduce the number of features and compute only relevant, which influence the income rate. In this experiment, we will apply the Boruta algorithm, where the features are copied, and then each new feature is filled randomly by shuffling its values. Random Forest is started on the resulting sample. In order to obtain statistically significant results, this procedure is repeated several times, variables are generated independently at each iteration.

On the third step, we apply the prediction algorithm to dataset except for irrelevant features. In the experiment we tested 3 methods, there are Gradient Boosting Regression, Random Fores Regressor, Ridge regression, Average method. During experiment we found out the tree-like structure algorithms is a set of decision's tree, which provide the best result of the prediction. Given a traning set $X = x_1, ..., x_n$ with responses income for each shop $Y = y_1, ..., y_n$, where N the count of shops. The Random Forest describes as:

$$\hat{Fr} = \frac{1}{B}\sum_{b=1}^{B} fr_b(x^{'})$$

where B is the number of trees and predictions for unseen samples $x^{'}$ can be made by averaging the predictions from all the individual regression trees on $x^{'}$.

The Gradient Boosting builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function [6].

$$\hat{Fg} = argmin\mathbb{E}_{x,y}[L(y, F(x))]$$

where $L(y, F(x))$ is loss function, which describes $L(y, F(x)) = (F(x) - y)$ so $\frac{-\partial L}{\partial F(x)} = 2(y - F(x))$. The parametr α has choosen 0.9, maximal depth = 3,the number of boosting stages to perform = 100.

However, our idea was the join in two Bagging and Boosting algorithms together that in this paper introduced as Average method. Indeed, we average the prediction result after test sample of two methods: Gradient Boosting Regression and Random Forest Regression.

$$\hat{Fav_i} = \frac{\hat{Fg_i} + \hat{Fr_i}}{2}, i = 0...22$$

The Regression methods calculate the value of prediction more inaccuracy, it proves the MAE and MAPE estimation in the fourth sections. Those algorithms are the basic tools of the statistics and them implementation was introduced in the paper [4]. For this reason, we did not introduce more information about them.

In this section we introduced the general prediction algorithms in machine learning. Also, we described the theoretical part of the retail location prediction module. On the next section we will focus on the retail store's features.

3 The retail features introduction

At the beginning, we want to consider the retail stores features and their influence on a store income rate. At the end of this section, we will introduce the features correlation heatmap.

Nowadays, the feature selection has been proccesed by general machine learning algorithms. They combine the computational complexity, the quick operation and universal, consequently, the choice of optimal algorithm depends on a formulation of the problem and the initial dataset. There are a few papers, which describe the retail feature analysis [2, 3]. One approach focuses on prediction of the retail location by using neural network [1]. We presume that an accuracy of prediction depends on a set of features, which influence income rate.

Indeed, each store has a number of features, which are usually common to all stores belonging to the same retail network. Thus, each feature influences the income rate with a greater or lesser degree, accordingly, the owner may manage to income by establishing the necessary measures of the features. Actually, there are no specific features that clearly affect a profit, because the number of important features depends on the category of the store network and the city urbanization. For example, in the case of pharmacy retail, an important feature is the presence of hospitals and medical institutions close to the store. We note that there are common features for all store chains, such as the area of the premises, the number of cash desks and sellers, the availability of parking, traffic, bus stations and others.

The paper will investigate the pharmacy industry, where the retail chain has the set of stores in the count of 22 items and each store has features in the count of 134 items.

Actually, the pharmacy industry has several specific types of the features. Some of them have a radially distance dependence on the store location. An example might be the count of bus station from the store at a distance of 100 meters, 200 meters, 300 meters, 400 meters, 500 meters and 800 meters radially. Another features is square of the store in the square meters. Table 1 describes a few features of the pharmacy industry.

Moreover, the heatmap of correlation matrix on a Figure 2 shows the dependence between several features and income. The deep red color identifies a strong positive correlation, but the deep-blue is a negative correlation. For instance, the feature "rubric 365-800" means a number of the hospital located within a radius of 800 meters has a strong correlation with the feature "pharmacies-800" - the number of competition on radius 800 meters.

In the next section, we will refer to the implementation of the prediction model on the pharmacy retail data set.

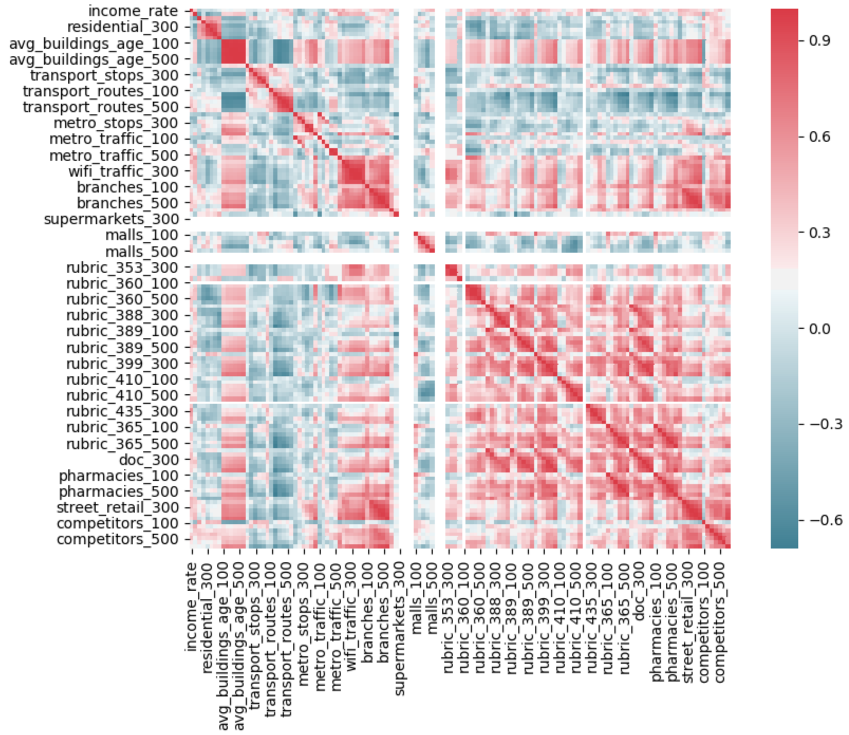


Figure 2: The correlation matrix of the retail features

4 The module prediction on an experimental retail data

The prediction module consists of the following steps,

1. The retail data preprocessing includes the scaling procedure. In brief, we use the min-maxscaling, which consists in rescaling the range of features to scale in the unit interval. In addition, the profitable to ranking the stores on income rate, where the first shop the low income and last the high income, to observe the income fluctuation.

2. Indeed, several features did not corelate to income and the count of featuers are massive that uninformative. For this reason, we need to select informative features, which correlated to income. Therefore, in this experiment, we apply two-step of selection, where on the first, the decision maker selects irrelevant features or noisy dataset. In the second, we apply feature selection algorithm Boruta, it works as a wrapper algorithm around Random Forest.

3. Next, we apply the prediction Average method to dataset after preprocessing procedure. Indeed, in the comparison diagram, we may observe the advantage of this method.

Table 1: The retail data before and after scaling procedure

storeID	area/scaling	residental 100/scaling	metro stops 200/scaling	malls 400/scaling
1	79.6/0.4688	331/0.4203	1/0.5	0/0
2	93.9/0.5750	513/0.7046	2/1	1/1
3	70.1/0.3982	432/0.5781	1/0.5	0/0
4	77/0.4495	293/0.3609	0/0	1/1
5	80.4/0.4747	702/1	0/0	1/1
22	101.7/0.633	288/0.3531	0/0	1/1

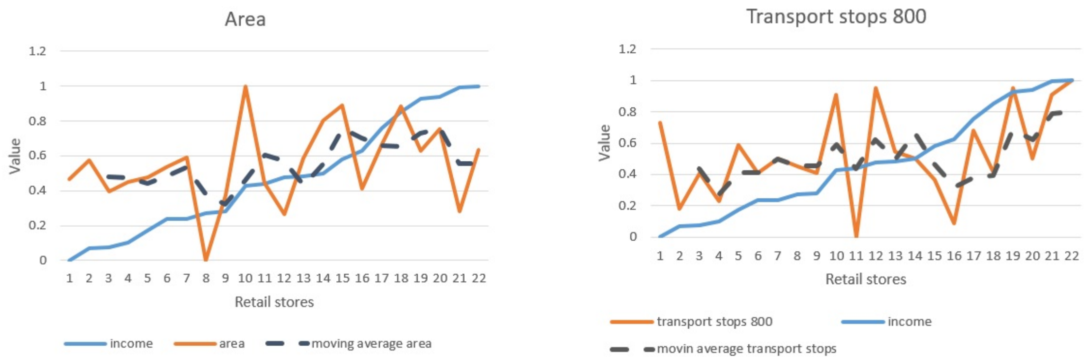
4. In the end, we build the project as a general module for income prediction. After, this module will integrate to the main business logic for the retail network system.

Certainly, before the avareged method implementation, we need to organize the retail store data preprocessing procedure. The experimental sample has 22 stores and 134 features. The experiment was implemented by using a Python 3.4 version. The dataset was given by the medicine company and was supplemented by Yandex services.

At the first, we need to structure information by ranking the stores according to the income rate. At the second, we implement the preprocessing procedure as the scaling of data set. It was computing by using min-max approach, which provides distribution values on the unit interval. The intermediate calculation was presented in Table 1.

After, we apply the Boruta algorithm to normalise dataset. Boruta is heuristic algorithm for selection of significant features based on the use of Random Forest. It tries to capture all the significant features in the dataset with respect to an outcome variable in our case is income. Therefore we defined 14 features, which depend on income change. There are "area", "transport stops 800", "metro traffic 400-500" and "metro traffic 800", "wifi traffic 100", "street retail 200 - 300" and etc. Indeed, Figure 3 describes two features "area" and "transport stops 800". On the first figure we observe the dependence on store's area and them income in the interval between 0.3 and 0.8. The moving average line shows the trend line for feature "area". The tails of moving the average line in interval [0;0.3] and [0.8;1] describes the low dependency on the store's income. The second figure shows the simillar case, wherein the interval [0.4;0.7] we observe the strong dependence on numbers of transport stations to radius 800 meters to the store's income.

According to this number of features, we apply the prediction algorithms to the preprocessing dataset. From a large number of prediction algorithms, we have choosen the Gradient Boosting Regression, Random Forest Regression and Avarage method. The tree structure algorithms provide more simple interpretation, flexibility, and high computational speed the specific dataset regardless of the count data and the variety. The results of the data prediction was provided by Table 2. Statistical estimates



a) Feature "area"

b) Feature "transport stops 800"

Figure 3: The dependence between income and features "area", "transport stops 800"

Table 2: Comparison the estimations of prediction algorithms

	MAE	MAPE
Gradient Boosting Regressor	0.29663477	142.6556371
Random Forest Regressor	0.261513839	123.5678737
Ridge	0.567570237	220.6778885
Average method	0.273655219	131.8967422

are used to estimate the prediction, where the simplest is the deviation of the fact from the prediction in quantitative terms. The value of mean absolute error supports Random Forest Regressor.

By the way, the dataset was devided on traning set and testing, but the number of samples was so small and was difficult for algorithms predict the exact result. So, we trained algorithms except for one store and we tested only this store. Other words, all stores are 22 iterations, where we except one store trained by 21 stores, where one was testing. During all iterations, we excepted each store from learning. Figure 4 shows the three prediction methods, at first thought all methods are simillar but the curve of Random Forest Regressor more flat than other and it MAE value is low. However, it is a fact, that the number of samples was poor and the algorithms do not provide exact result of the prediction, because they necessary the much more quantity of sample. This problem we may observe on the tails of curve, where the low-income stores has only 4 units and a similar situation with the high-income stores.

In summery, we computed all steps for explaining the Prediction module which was included on the union project for the pharmacy retail store's network. By using this program common person may analyze the location on a city map according to the features and follow to this prediction's advice.

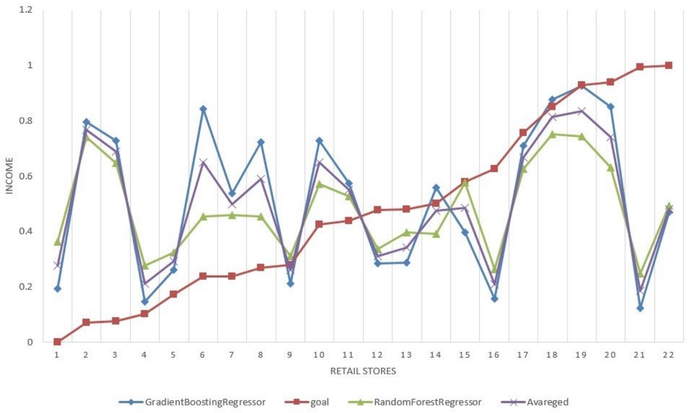


Figure 4: The comparative diagram of prediction store's network income

5 Conclusion

In this paper, we focused on a pressing task as a prediction income of the retail chain. The prediction module based on Average model, which is union the Gradient Boosting Regression and Random Forest methods, the feature selection was computed by Boruta method. Currently, this module would be extended not only the pharmacy industry but also on the other business area. Despite it, this module was included in the global service "Mestomer".

References

[1] Satman M.H., Altunbey M. (2014). Selecting Location of Retail Stores Using Artificial Neural Networks and Google Places API. *International Journal of Statistics and Probability*. Vol. **3**, pp. 67-77.

[2] Ferreira K., Lee B.H., Simchi-Levi D. (2016). Analytics for an Online Retailer: Demand Forecasting and Price Optimization. *Manufacturing & Service Operations Management*. Vol. **18**, pp. 69-88.

[3] Glaeser C.K., Fisher M., Su X. (2016). Optimal Retail Location: Empirical Methodology and Application to Practice. *SSRN Electronic Journal*. pp. 1-28.

[4] Duda R.O., Hart P.E., Stork D.G. (2000). *Pattern Classification (2Nd Edition)*. Wiley-Interscience, New York, NY, USA.

[5] Cheng E., Li H., Yu L. (2007). A GIS approach to shopping mall location selection. *Building and Environment*. Vol. **42**, pp. 884-892.

[6] Friedman J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*. Vol. **29**.