Zhang Qigong

gsangeryeee@gmail.com

5 January 2018

Store Sales Forecast

I. Definition

Project Overview

Sales forecast is an integral part of JD.com's credit evaluation for small business loan. Accurate sales prediction translates into loan efficiency, improving capital utilization.

Background

JD Finance's supply chain finance has over 100,000 enterprise clients, providing USD 250 billion of loans in total with focus on serving SMEs. Supply chain finance is an industry supply chain based financial activities, its purpose is through the industrial supply chain to provide support for financial activities. In the past few years, it launched products like Jing Bao Bei, Jing Xiao Dai, and Dong Chan Rong Zi based on JD's big data. Among them, Jing Xiao Dai is a credit-based financial product, with high degree of loan autonomy, zero collateral, instant grants, low-cost financing, no payback terms, and online approval procedure.

Enterprises can log into the JD financial platform using store accounts for details on loan qualifications and application. Once successful, loans will be put into the store account, seamlessly connecting the payment procedure with JD. To make fintech even more beneficial, JD supply chain finance incorporates emerging companies with tradi-

tional companies, providing embedded service and technical support for online and offline risk control.

Details

Measuring and tracking store loan is the key to forecasting store sales. Through forecasting sales, we could better assess the financing needs and loan amount for each shop. This task requires the team to establish a forecasting model base on the simulated data of sales records, merchandise information, product evaluation, and advertising costs of the stores provided by the contest to predict the sales of the stores in the coming 90 days.

Datasets are simulated data based on real business scenario. All the data are sampled and desensitized. Field values and distributions are different from the real business data.

The output variable (sales of the stores) takes continuous values. So the first step, this is a regression problem.

For this type of problem, I will use some of regression tools. Such as Linear Regression, Support Vector Machines, K-Nearest Neighbor Regression, Decision Trees and Ensemble methods.

In general, I will try to solve this problem in four steps.

1. Pre-processing
2. Data processing
3. Evaluate and choose Algorithms
4. Optimize the selected algorithm

Metrics

I use mean absolute error (MAE) to measure performance of a model. Because of the problems to be solved, I need to evaluate the error between the true value an the predicted value so I have to choose between mean absolute error(MAE) and mean squared error (MSE). For this problem, I am not concerned about large errors whose consequences are not much bigger than equivalent smaller ones. The MAE is more robust to outliers since it does not make use of square. So I choose the MAE as performance metrics.

For each shop, to estimate the difference between the real sale revenue and predicted revenue, scores will be calculated based on the following formula: yi is the real value, y_hati is the predicted value, and m represents the total number of shops to be evaluated.

II. Analysis

Data Exploration

1.TRAINING DATASET

Training dataset includes daily orders, sales number, number of customers, number of reviews, advertising costs, and etc. within 270 days prior to 2017-04-30; off-shelf merchandise after 2017-04-30; on-shelf merchandise; and shop sales number from 2016-06 to 2017-01.

2.TESTING DATASET

Teams are required to forecast total sales within 90 days of 2017-04-30 for each store (for all shop ids involved in the training dataset).

3.FILE INFORMATION

File Name	Data

t_order.csv	Order information
t_product.csv	Product information
t_comment.csv	Reviews
t_ads.csv	Advertising information
t_sales_sum.csv	Sale revenue of last 90 days from each month

4.DICTIONARY

File name	Field name	Field Description
t_order	shop_id	shop id
	pid	product id
	ord_dt	order date
	ord_cnt	order number
	sale_amt	sales amount
	user_cnt	number of customers
	rtn_amt	total refund
	rtn_cnt	number of refund order
	offer_amt	total discount
	offer_cnt	number of discount
t_product	shop_id	shop id
	pid	product id
	brand	brand id
	cate	category id
	on_dt	date on shelf
	off_dt	date off shelf
t_comment	shop_id	shop id
	bad_num	number of negative reviews
	mid_num	number of neutral reviews
	good_num	number of positive reviews
	dis_num	number of unboxing reviews
	cmmt_num	number of total reviews
	create_dt	date to create
t_ads	shop_id	shop id
	charge	advertising fee refill

t_ads	consume	advertising spending
	create_dt	date created
	shop_id	shop id
t_sales_sum	sale_amt_3m	sale revenue in the next 90 days of each month end
	dt	last day of each month

Looking at the original datasets, I found that:

1. There was duplicate data in t_sales_sum.csv.
2. There are NULL values in datasets.
3. Information for merchandise pulled off the shelf before 2017-04-30 are not provided for practical considerations,

FIle name	number of record	features
t_ads.csv	226128	3
t_comment.csv	636206	6
t_order.csv	12098397	9
t_product.csv	11398239	5
t_sales_sum.csv	24000	1

You can download data from here http://jddjr.joybuy.com/item/10

Exploratory Visualization

The plot below shows the changes in the values(sale revenue in the next 90 days of each month end) over a period of eight months.

From the plot we can see that the majority of stores didn't have upward or downward trends.

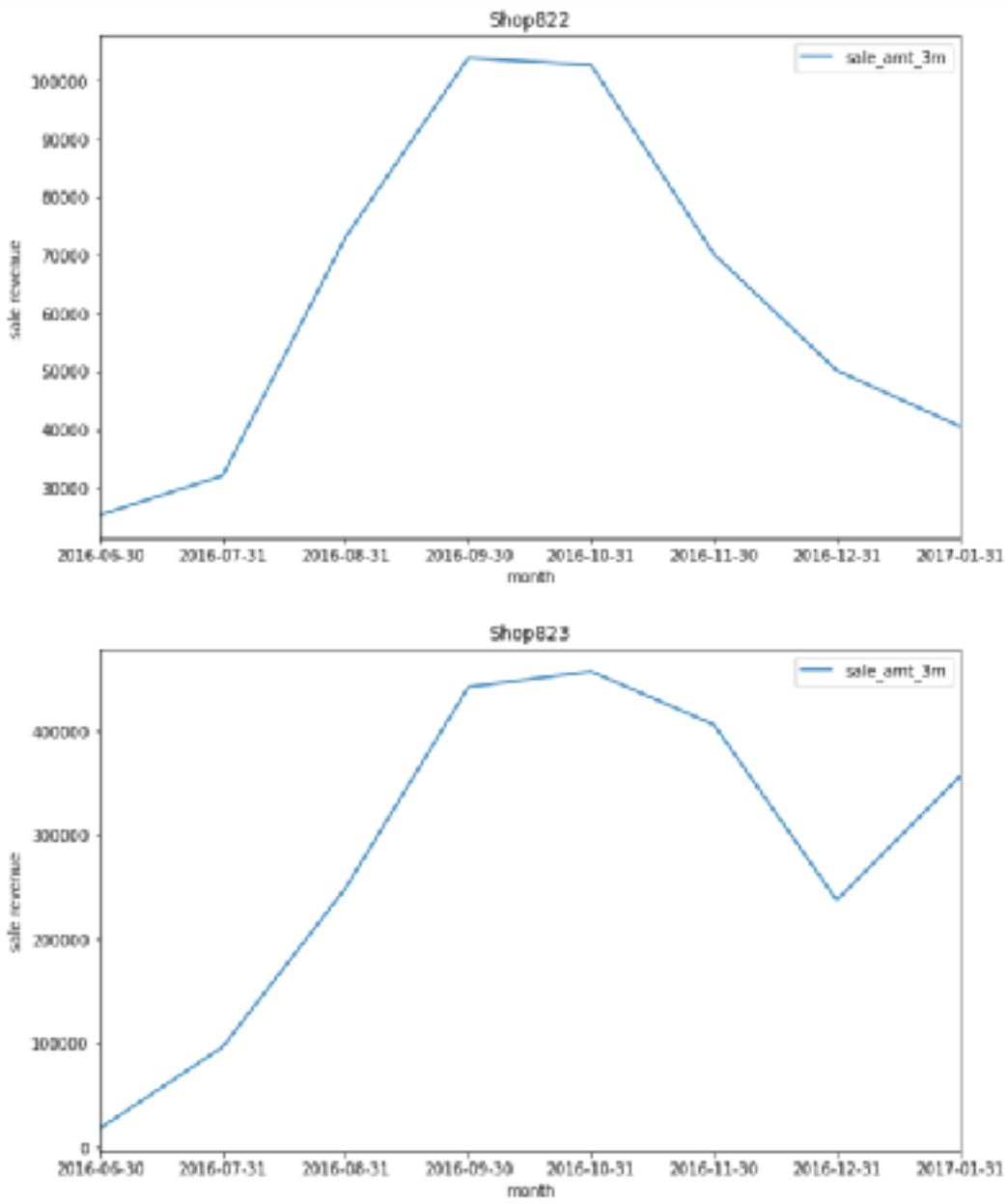


Figure 1. Sample of different stores sale revenue in the next 90 days of each month end in 8 months.

Algorithms and Techniques

Sale revenue is a continuous variables, so sales forecasting is a regression prediction problem. There are several models such as Linear Regression, Support Vector Machines, K-Nearest Neighbor Regression, Decision Trees and Ensemble methods.

From the Sciklit-learn toolkit model usage advice. I will try Linear Regression, Support Vector Regression, K-Nearest Neighbor Regression and Ensemble methods.

Because Linear regression in the most simple and easy to use regression model. Without knowing the relationship between the features, we can use the linear regression model as the baseline system. SVR performs well on small amounts of data and high dimensional predictions, so I choose SVR and using different kernel. Because of poor generalization of Decision Trees model, the stability in not high, so I did not choose this model. K-Nearest Neighbor Regression and Ensemble methods have good performance on the regression problem, so I will add these two model to solve this problem.

First of all , I'll compare several models and choose the best one that It is the best at MAE evaluation .After I'll be tuning the parameters in order to get better performance.

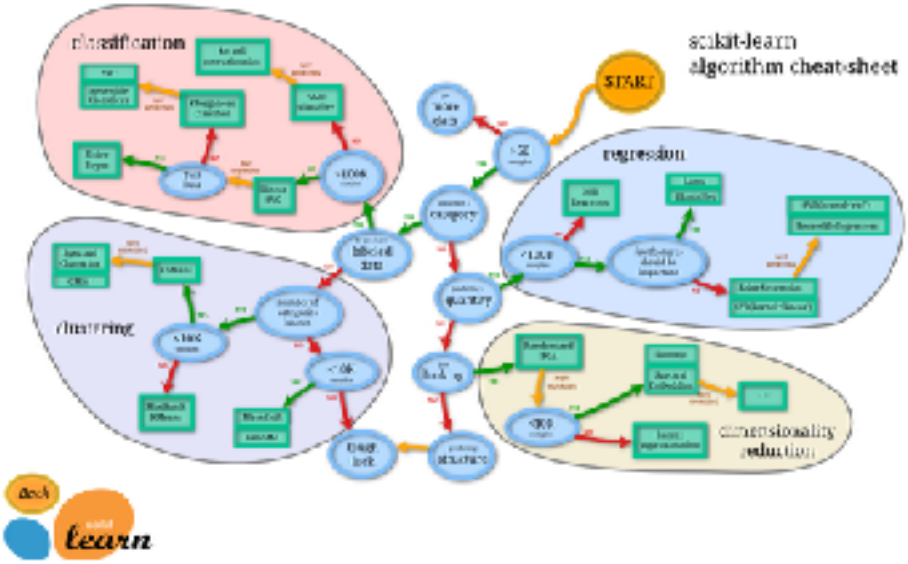


Figure 2. Scikit-learn toolkit model usage advice (http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html)

I will describe mainly models principle.

Linear Regression：

1. Linear Regression hypotheses function.

$$h(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x,$$ (hypotheses h)

2. Linear Regression Cost Function

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2.$$

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta).$$

3. The update rules are as follows
4. In the process of solving the optimal value in the linear regression model, aa least squares algorithm is used to estimate the parameters. In linear regression, least squares are trying to find a straight line, and the sum of all the Euclidean distances of all the samples to the straight line is the smallest. Let's take a look at how to find the value of h when the minimum value of J is obtained. The value of theta is also obtained indirectly.
5. To solve this, we use a gradient descent algorithm to describe this process. Select the initial w, and then continue learning to update w, to achieve the optimal solution

Support Vector Regression

1. Use a function to fit the relationship between x and y. For SVR, x is a vector and y is a scalar. The function of the fit is y = W ^ T * g (x) + b, where g (x) is the feature space vector corresponding to the kernel.
2. The SVR believes that as long as the estimated y is within a constant epsilon on either side of the actual y, it is assumed that the estimate is correct without any loss

3. The goal of SVR optimization is | W | min, so that the y-x curve has the smallest slope and the function is the flattest, which is said to increase the robustness of the estimation.
4. After things there can be soft margin, with a small positive control. Use dual to solve
5. Given training vectors and a vector y, SVR solves the following primal problem:

$$\min_{w,b,\zeta,\zeta^*} \frac{1}{2} w^T w + C \sum_{i=1}^{n} (\zeta_i + \zeta_i^*)$$

$$\text{subject to } y_i - w^T \phi(x_i) - b \leq \varepsilon + \zeta_i,$$
$$w^T \phi(x_i) + b - y_i \leq \varepsilon + \zeta_i^*,$$
$$\zeta_i, \zeta_i^* \geq 0, i = 1, \ldots, n$$

6. Its dual is

$$\min_{\alpha,\alpha^*} \frac{1}{2} (\alpha - \alpha^*)^T Q (\alpha - \alpha^*) + \varepsilon e^T (\alpha + \alpha^*) - y^T (\alpha - \alpha^*)$$

$$\text{subject to } e^T (\alpha - \alpha^*) = 0$$
$$0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, \ldots, n$$

   where e is the vector of all ones, C>0 is the upper bound,Q is an n by n positive semidefinite matrix.

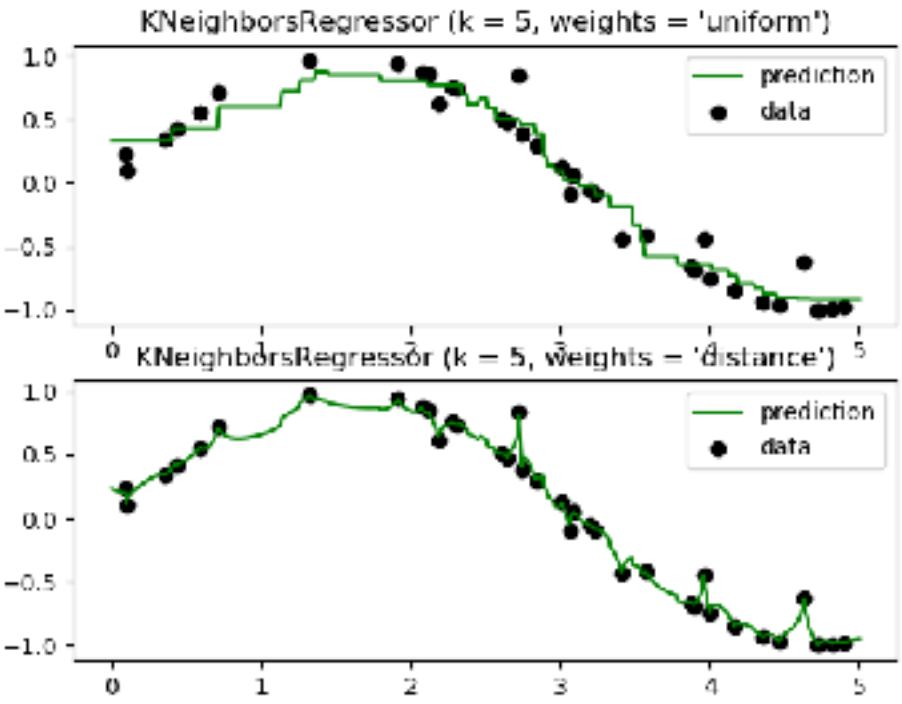
7. $Q_{ij} \equiv K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is the kernel.

KNeighborsRegressor:

1. The basic nearest neighbors regression uses uniform weights: that is, each point in the local neighborhood contributes uniformly to the classification of a query point.
2. Under some circumstances, it can be advantageous to weight points such that nearby points contribute more to the regression than faraway points. This can be accomplished through the weights keyword.
3. The default value, weights = 'uniform', assigns equal weights to all points. weights = 'distance' assigns weights proportional to the inverse of the distance from the query

point. Alternatively, a user-defined function of the distance can be supplied, which will be used to compute the weights.

4. An example of using KNN fitting process is shown in the figure:



Benchmark

For simplicity, I decided to use the mean of 8 months sale_amt_3m(t_sales_sum.csv) in each store as the benchmark.

III. Methodology

Data Preprocessing

1. Remove duplicate data. In t_sales_sum.csv have duplicate data need to be removed.

2. I chose 2016-09,2016-10,2016-11,2016-12,2017-01 these five months data as training datasets, because I found that these five months' data is complete in each original file.
3. I chose 2017-04 data as testing datasets.
4. In order to facilitate teh calculation, I separately preprocess teh data in each original file, and then generate a training datasets and testing datasets.
5. Each shop in accordance with the above five training time and test time period,

* t_order: Statistics during the above period the total of sale amount, total refund, total discount, total order number, total number of customers.

* t_proudct: Statistics during the above period the total of the products on the shelf.

* t_comment:Statistics during the above period the total amount of unboxing reviews and positive reviews rate and negative reviews rate.

* t_ads: Statistics during the above period the total amount of advertising fee refill and the total advertising spending.

* For missing data, I use median and mean to fill.

Implementation

1. In time series predictions, I chose 3 month (2016-09 to 2016 -11)datasets as training data, 1 month (2016-12) datasets as validation data and 1 month (2017-01) datasets as testing data.
2. In the initial inspection of the data, I found that there was a big difference between the predicted sale revenues. So we first need to standard scale the features and target values.
3. Using Linear regression, Support Vector Regression(using linear and rbf

kernels),KNeighborsRegressor,DecisionTreeRegressor and Ensemble methods to learn store training datasets and predict the test data.

4. Using R-squared, mean squared error and mean absolute error to evaluate the results.

5. By comparing the scores, I found that the Support Vector Regression with rbf kernel gets the highest sores in mean absolute error, so I chose this method to predict the target.

Model	MAE
LinearRegression	86454.21
SVR(kernel='linear')	81585.17
SVR(kernel='rbf')	79493.08
KNeighborsRegressor(weights = 'uniform')	89645.16
KNeighborsRegressor(weights = 'distance')	89619.63
RandomForestRegressor	95157.02
ExtraTreeRegressor	95008.81
GradientBoostingRegressor	86041.86

6. The finial output to the csv file.

Refinement

I have trained a model using SVR with RDF kernel. To ensure that I am producing an optimized model, I will train the model using the grid search technique to optimize the parameter for the SVR.

In SVR, the important parameters are gamma and C. The parameter C is penalty parameter C of the error term. If the C is higher, the model is easy to overfit. The smaller C is easy to underfit.

The Gamma is a parameter that comes with the function when the RBF function is selected as the kernel. The larger the gamma, the fewer the support vectors, the smaller the gamma, the more support vectors. The number of support vectors affects the speed of training and prediction.

gamma	C	MAE
0.001	0.001	104764.28
0.001	0.01	94783.19
0.001	0.1	81586.96
0.001	1	80095.25
0.001	10	79369.51
0.01	0.001	94587.63
0.01	0.01	81460.72
0.01	0.1	79504.74
0.01	1	78587.76
0.01	10	78369.52
0.1	0.001	88860.92
0.1	0.01	81021.37
0.1	0.1	78940.50
0.1	1	78418.21
0.1	10	78687.83
1	0.001	96514.49
1	0.01	88693.65
1	0.1	83464.75
1	1	83066.76
1	10	88160.89
0.05	1	79043.26

As result,Parameter 'C' is 1 for the optimal model and parameter 'gamma' is 0.05 for the optimal model

IV. Results

Model Evaluation and Validation

Finally, I use the SVR with RDF kernel and use these parameters to make predictions. After several sets of performance test, I found that the models under different configurations have very large performance differences on the same test set. And using the Radial basis function kernel function to no linearly map the features, SVR shows the best regression performance.

For validation, I use the validation set (2016-12-31) by final model. I got MAE score that is 75854.46 better than training data score (79043.26)

The final predict file is ‘Sales_Forecast_Upload_3_opt'.

Justification

Overall, this model is useful for most stores to forecast sale revenue in the next 90 days of each month end. Model for such sales forecasting problem is universal, versatility. For this project, the model can predict sales more accurately and translates into loan efficiency.

As shown in the following table, we take the average of the “shop1” and the sales calculated using the model to calculate the MAE values respectively. It can be seen from the comparison that the results obtained by using the model are better than the benchmark.

Shop 1	sale revenue	MAE Score
the mean of 8 months	272555.11	53457.27
SVR(kernel='rbf')	269203.39	50105.55
True value	219097.84	

V. Conclusion

Free-Form Visualization

The 2017-04-30 forecast data added to the t_sales_sum.csv file, re-create each store's 9-month sales figure 3.

As you can see from the figure, the trend of the image is coherent after adding the 2017-04-30 forecast.

Reflection

The difficulty of solving this problem lies in how to find a model that can not only predict the sales more accurately but also have some commonality that can be used in most shops. On the surface, the issue is time-related, but not seasonal. Because the focus of the problem is to guide the company's review and payment of the loan to the store through the prediction of the sales of the store. So I'm solving the problem is not specifically for each shop or seasonal cycle to carry out training. I am trying to find a more generic one that has nothing to do with the specific store and time.

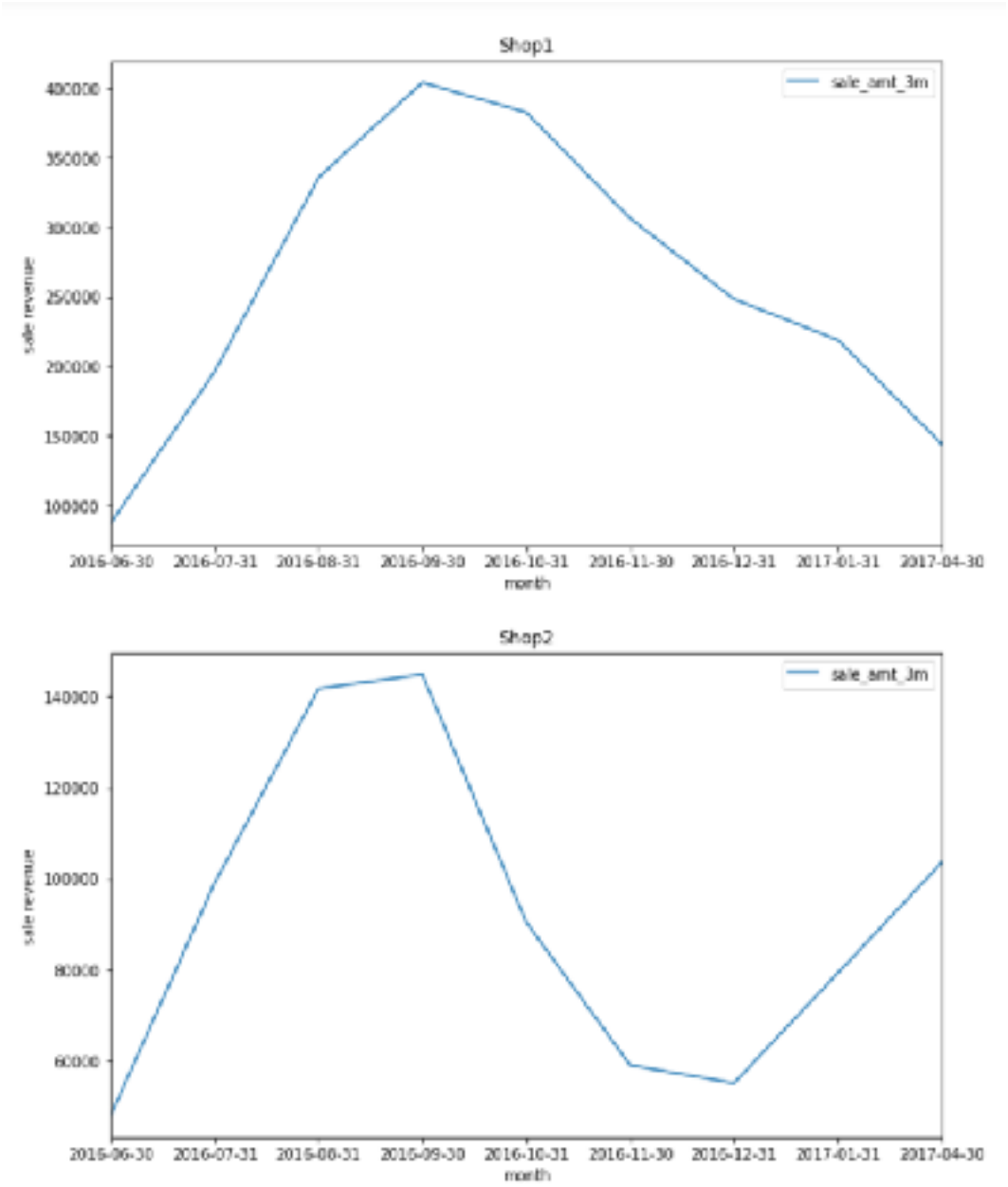


Figure 3. Sample of different stores sale revenue in the next 90 days of each month end in 9 months.

Improvement

In solving this problem, I was thinking that the sales of the shops will increase for special holidays and the demand for loans will also increase. This model does not consider this factor for a while, and further I can participate in the training by adding the feature of festival.