FLIP00 FINAL PRESETANTION REPORT

RONGXIN XU

Abstract. This report contains five parts. First, introduce the definition of the problem, describe the data and analyze the problem. Second, statistical the data information, visualize the data to find some potential relationships between the attribute values and process the data like dummy variables, feature engineering, feature selection etc. Third, explain the most important parameters of different algorithms, and the method of experiment. Fourth, experiment and analyze the performance of different algorithms based on experimental result. The last one is conclusion .

Contents

Date: 2019-11-29.
1991 *Mathematics Subject Classification.* Artificial Intelligence.
Key words and phrases. Machine Learning, Data Mining, ...

1. Introduction

1.1. **Problem Statement.** This is a problem with time-series prediction. After a month of making scientific observations and taking careful measurements, can predict total sales for every product and store in the next month. The raw dataset contains train set with 2935849 samples and 214200 unlabeled samples as test set. Through the train data, predict total sales for every product and store in the next month.

1.2. **Data List.**

There are 6 data sets with a total of 11 attributes, the flowings are the name and meaning of attributes.

id: an Id that represents a (Shop, Item) tuple within the test set.
shop_id: unique identifier of a shop.
item_id: unique identifier of a product.
item_category_id: unique identifier of item category.
item_cnt_day: percentage of soul in the creature.
item_price: current price of an item.
date: date in format dd/mm/yyyy.
date_block_num: unique identifier of item category.
item_name: name of item.
shop_name: name of shop.
item_category_name: name of item category.

1.3. **Problem Analysis.**

1.3.1. *Problem Possible Solutions.*

There are many machine learning algorithms can solve the Time series prediction problem, such as xgboost, random forest and so on. Use CV to find the best parameters of the algorithms and then validate with testing data. But the most important thing is do feature engineering to improve accuracy.

1.3.2. *Evaluation Methods.* Before experiment, determine the evaluation methods to assess the model performance is very important, usually it has the following methods for classification problem:

- RMSE

2. Exploratory Data Analysis

2.1. **Data Information.**

The following table 1 is the statistical result of each attribute in sales_train.csv. There are 6 numerical variables, and no missing values. The data is very clean and complete, So let's start visual analysis.

2.2. **Data Visualization.**

Use EDA to plot the distribution of the data, can observate the data intuitively and find the relation between the attribute values. For example boxplot can visually observe the distribution of numerical variables, scatterplot can show their distribution trends and whether exists outliers. For classification problems, the data with the same label is drawn in same color, which is very helpful for the construction of the Feature.

Table 1. Data Information

	date_block_num	shop_id	item_id	item_price	item_cnt_day	item_category_id
count	2935849	2935849	2935849	2935849	2935849	2935849
mean	14.57	33	10197.23	890.62	1.24	40
std	9.42	16.23	6324.3	1726.44	2.62	17.1
min	0	0	0	-1	-22	0
25%	7	22	4476	249	1	28
50%	14	31	9343	399	1	40
75%	23	47	15684	999	1	55
max	33	59	22169	307980	2169	83

2.2.1. *Histogram.*

The figure Figure 1 shows the distribution of the various attributes. It seems that item_id and shop_id has a huge impact on sales and sales tend to decline with the date.

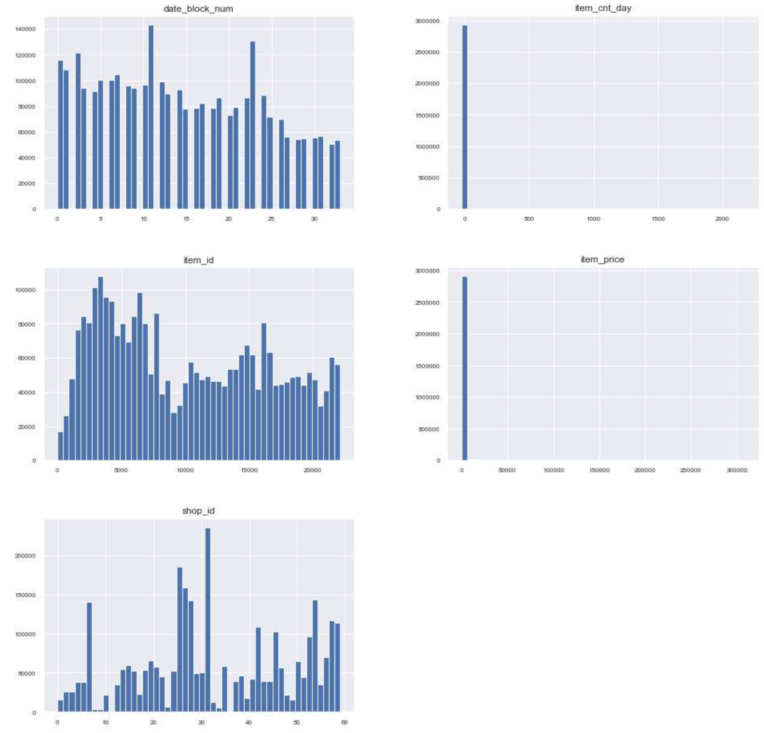

Figure 1. Distribution of individual variables

2.2.2. *Boxplot.*

When analyzing the data, the boxplot can effectively help us identify the characteristics of the data: visually identify outliers in the dataset or determine the data dispersion and bias of the data set. Through the figure Figure 2, we know that the outliers are very small, so can be ignored.

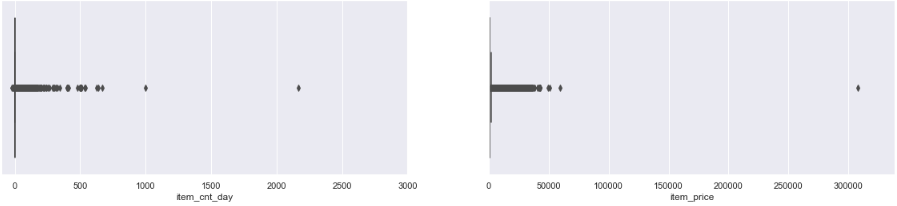

Figure 2. Boxplot of item˙cnt˙day and item˙price

2.2.3. *Scatterplot Plot.*

Pairwise plot is a favorite in exploratory analysis to understand the relationship between all possible pairs of numeric variables. This pairplot Figure 3 shows that data is distributed normally. And while most pairs are widely scattered (in relationship to the type), some of them show clusters: hair_length and has_soul, hair_length and bone_length. So it may need to reassemble the data.

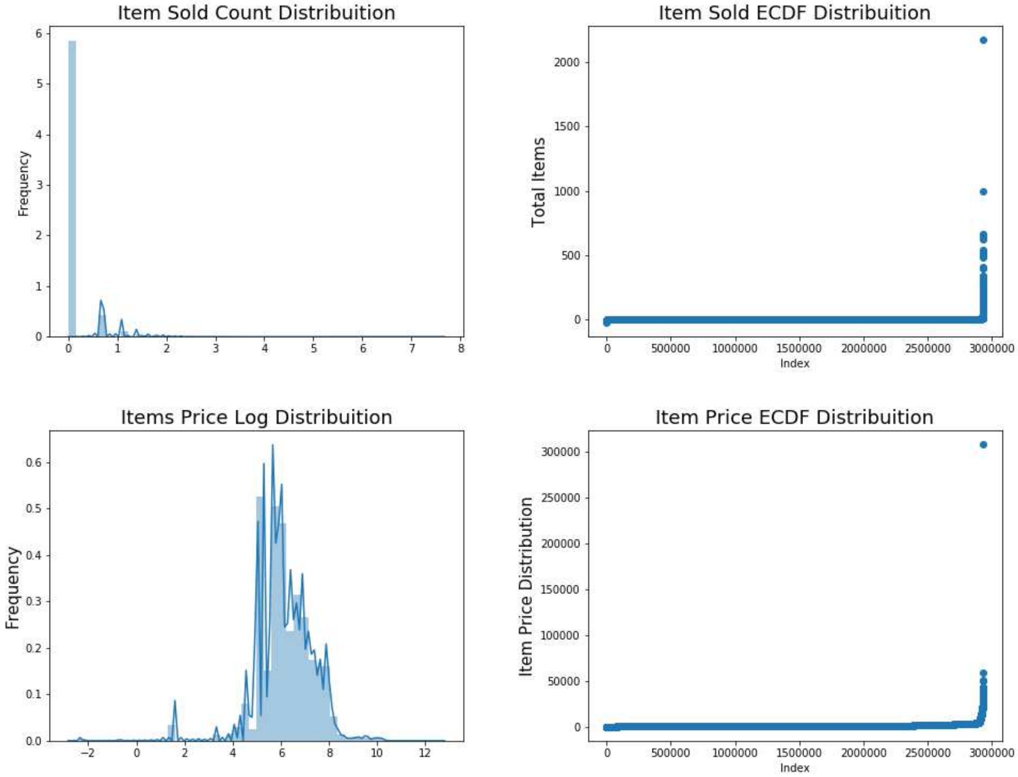

Figure 3. Scatterplot of item˙cnt˙day and item˙price

2.2.4. *Correllogram.*

Correlogram is used to visually see the correlation metric between all possible pairs of numeric variables in a given dataframe. This figure Figure 4 make it convenient for us to analyze features. You can see that the item˙cnt˙day related to the target to be analyzed is item˙id and item˙price.

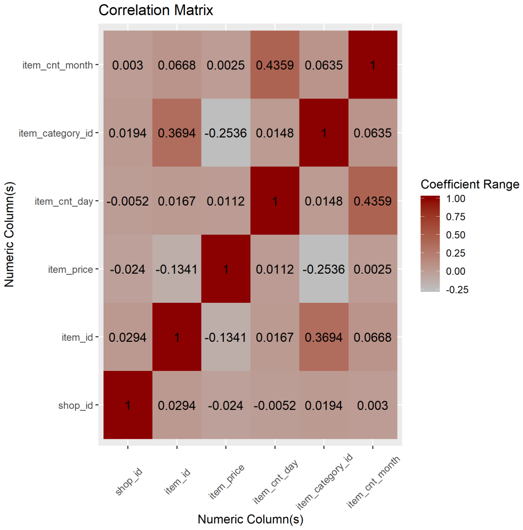

FIGURE 4. Correllogram

2.3. Data Preparation.

2.3.1. *Feature Selection.*

Use the algorithm below to calculate the importance of features. The following figure Figure 5 is a histogram ordered by feature importance.

Algorithm 1 Features Selection

```
Require: Features $X = \{X_1, X_2, ..., X_n\}$, The number of tree node $M$, $GI_m$ Gini index
    of node $m$, $K$ the number of target, $p_mk$ proportion of target $k$ in node $m$, $VIM_{jm}^{(Gini)}$
    the importance of feature $X_j$ in node $m$ , $n$ the tree number of RF.
Ensure: Variable Importance Measures $VIM_j^{(Gini)}$.
Initialize $GI_m$ , $VIM_j^{(Gini)}$;
for $m \leftarrow 1...M$ do
    for $k \leftarrow 1...K$ do
        Compute the Gini index of node $m$ $GI_m = \sum_{k=1}^{|K|} \sum_{k' \neq k} p_{mk} p_{mk'} = 1 - \sum_{k=1}^{|K|} p_{mk}^2$
    end for
    Divide node m into node r and node l
    Compute the importance of feature $X_j$ in node $m$ $VIM_{jm}^{(Gini)} = GI_m - GI_l - GI_r$
end for
for $i \leftarrow 1...N$ do
    Compute variable importance measures $VIM_j^{(Gini)} = VIM_j^{(Gini)} + VIM_{ij}^{(Gini)}$
end for
return $VIM_j^{(Gini)}$
```

We take these features to form a new train datad.

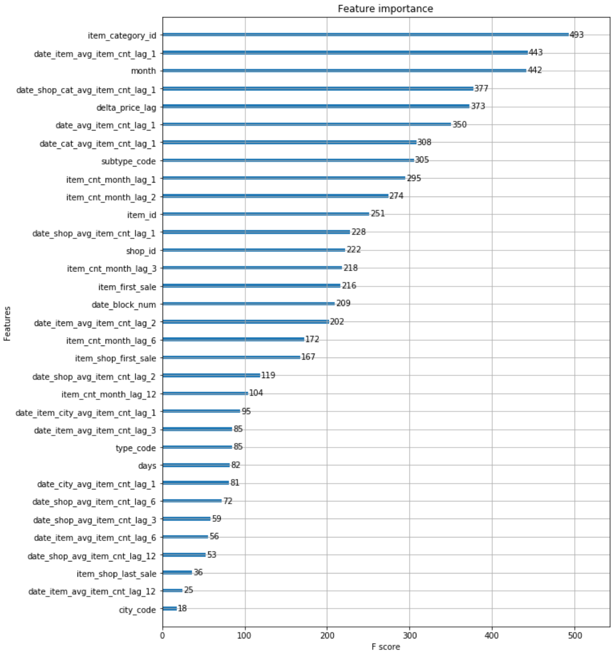

FIGURE 5. Feature Importance

3. METHODS

There are many machine learning algorithms for Time-series problem. Choose the following algorithms as the base models of ensemble model,show the most important parameters.

- Base Models
  - RandomForest
  - XGBoost
  - LSTM
  - Linear regression
  - KNN
- Ensemble Model

3.1. **Base Models.**

The base models have many parameters, select the some parameters that have a larger impact on the forecast results, the use Grid Search to find the optimal paratemers set. The following is training result.

3.1.1. *RandomForest.*

Random forest is a classifier with multiple decision trees, and the output is determined by the mode of the individual tree output.

n˙estimators: the number of decision trees
criteriom: criterion of choosing the most appropriate node
max˙depth: The maximum depth of the tree, the default is None
max˙features: The feature that is divided when selecting the optimal attribute cannot exceed this value.

3.1.2. *XGBoost.*

XGBoost is to establish K regression trees so that the predicted value of the tree group is as close as possible to the true value (accuracy) and has the greatest generalization ability. From a mathematical point of view, this is a functional optimization, multi-target.

learning˙rate: control the speed of each update
n˙estimators: number of iterations
max˙depth: the depth of tree
gamma: penalty factor
subsample: the proportion of data used in all training sets when training each tree
colsample˙bytree: the proportion of features used in all trees when training each tree

3.1.3. *LSTM.*

Long short-term memory (LSTM) is an artificial recurrent neural network (RNN) architecture used in the field of deep learning. LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video).

units: Output dimension the number of neurons in the i-th hidden layer
activation: activation function
recurrent˙activation: Activation function applied to the loop step
use˙bias: Boolean, whether to use bias term

3.1.4. *Linear regression.*

Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).

fit˙intercept: whether to calculate the intercept for this model. If set to False, no intercept will be used in calculations the number of neurons in the i-th hidden layer
normalize: This parameter is ignored when fit˙intercept is set to False
copy˙X: If True, X will be copied; else, it may be overwritten
n˙jobs: The number of jobs to use for the computation

3.1.5. *KNN.*

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification.

n˙neighbors: Number of neighbors to use by default for kneighbors queries
weights: weight function used in prediction. Possible values
algorithm: Algorithm used to compute the nearest neighbors
leaf˙size: Leaf size passed to BallTree or KDTree

3.2. **Ensemble Model.**

To combine the base models as 1st level model predictions, I'll use a simple linear regression. As I'm only feeding the model with predictions I don't need a complex model.

4. Experiment and Analysis

In the Data Exploration, I has created some new feaures and found some outliers. Because the number of outliers is very small, after ignoring the outliers, the features are selected for experiments based on their importance. And use the trained models as the base models of ensemble model, then do experiment.

4.1. Base Models Training Result.

The following are the best parameters and the best Score in training of the base models.

- Best Parameters of Models
  RandomForest: ’n˙jobs’: ’-1’, ’max˙depth’: 15, ’random˙state’: 42, ’n˙estimators’: 25
  XGBoost: ’max˙depth’:10, ’subsample’:1, ’min˙child˙weight’:0.5, ’eta’:0.3, ’num˙round’:1000, ’seed’:1, ’silent’:0, ’eval˙metric’:’rmse’
  LSTM: ”batch˙size”:128, ”verbose”:2, ”epochs”:10
  Linear regression: No other parameters required
  KNN: n˙neighbors=9, leaf˙size=13, n˙jobs=-1

4.2. Forecast Result of Base Models.

From the Table 2, it shows that the rmse of each model, RandomForest and XGBoost perform better, and there is not much difference between the other models.

Table 2. Best Score of the Base Models

	RandomForest	XGBoost	LSTM	Linear regression	KNN
Train rmse	0.8358	0.8327	0.9276	0.8572	0.6976
Validation rmse	0.8810	0.8959	0.6611	0.8806	0.8946

4.3. Forecast Result of Ensemble Model.

Ensemble model means using more than 1 model to finish the prediction. The train rmse is 0.764973649571408.

5. Conclusion

- Exploratory data analysis is very important for the competition, Discover the imperfections of the data and have a certain understanding of the overall appearance of the data, which will help later modeling and analysis.
- The data that we have, needed processed in many cases. Data preprocessing includes deal with missing data and outliers, We must think carefully about the outliers, such as ignoring them.
- The most important thing is feature engineering. We have to think carefully and deal with outliers, such as ignoring or deleting them.
- There is no best model, only the best model. We should try as many models as possible to get the best prediction results.
- Feature engineering is very important and even plays a decisive role in this competition.
- The Ensemble model may perform better than a single model when dealing with some complex problems.

References

List of Todos

(A. 1) School of Business Administration,, Hunan University, Changsha 410012, China
Email address, A. 1: `rongxin_xu@163.com`