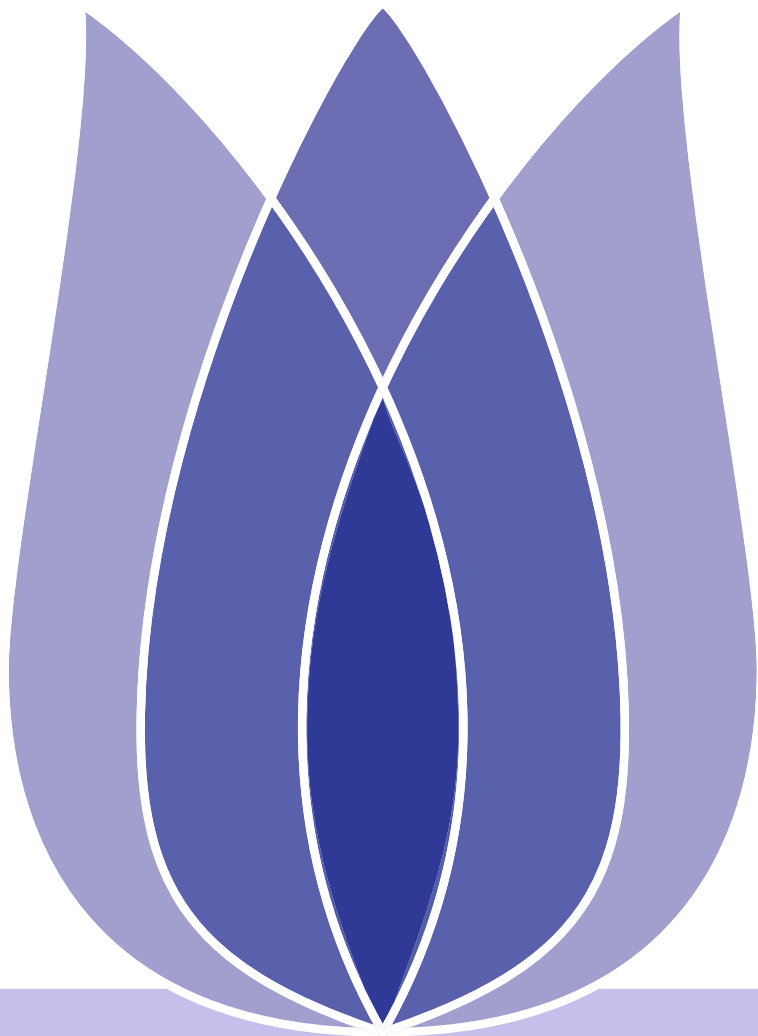


FLIP(00) Final-term Presentation

Rongxin Xu
Hunan University

29 November 2019





Outline

- [Problem Statement](#)
- [Exploratory Data Analysis](#)
- [Feature Engineering](#)
- [Methods](#)
- [Forecast Results](#)
- [Conclusion](#)

- Problem Statement**
- Exploratory Data Analysis**
- Feature Engineering**
- Methods**
- Forecast Results**
- Conclusion**



Problem Statement

Problem Description

Data Set

Exploratory Data Analysis

Feature Engineering

Methods

Forecast Results

Conclusion

Problem Statement



Problem Description

Problem Statement
Problem Description
Data Set
Exploratory Data Analysis
Feature Engineering
Methods
Forecast Results
Conclusion

Defn 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.



- Problem Statement
- Problem Description
- Data Set
- Exploratory Data Analysis
- Feature Engineering
- Methods
- Forecast Results
- Conclusion

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

■ Data List

- 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.



- [Problem Statement](#)
- [Exploratory Data Analysis](#)**
- [Data Information](#)
- [Data Visualization](#)
- [Data Visualization](#)
- [Data Visualization](#)
- [Data Visualization](#)
- [Feature Engineering](#)
- [Methods](#)
- [Forecast Results](#)
- [Conclusion](#)

Exploratory Data Analysis



Data Information

- Problem Statement
- Exploratory Data Analysis
- Data Information**
- Data Visualization
- Data Visualization
- Data Visualization
- Data Visualization
- Feature Engineering
- Methods
- Forecast Results
- Conclusion

The following 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.

	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



- [Problem Statement](#)
- [Exploratory Data Analysis](#)
- [Data Information](#)
- [Data Visualization](#)
- [Data Visualization](#)
- [Data Visualization](#)
- [Data Visualization](#)
- [Feature Engineering](#)
- [Methods](#)
- [Forecast Results](#)
- [Conclusion](#)

Exp Use EDA to plot the distribution of the data, can observate the data intuitively and find the relation between the attribute values.

- **Figures**
 - ◆ Histogram
 - ◆ Boxplot
 - ◆ Scatterplot Plot
 - ◆ Correllogram

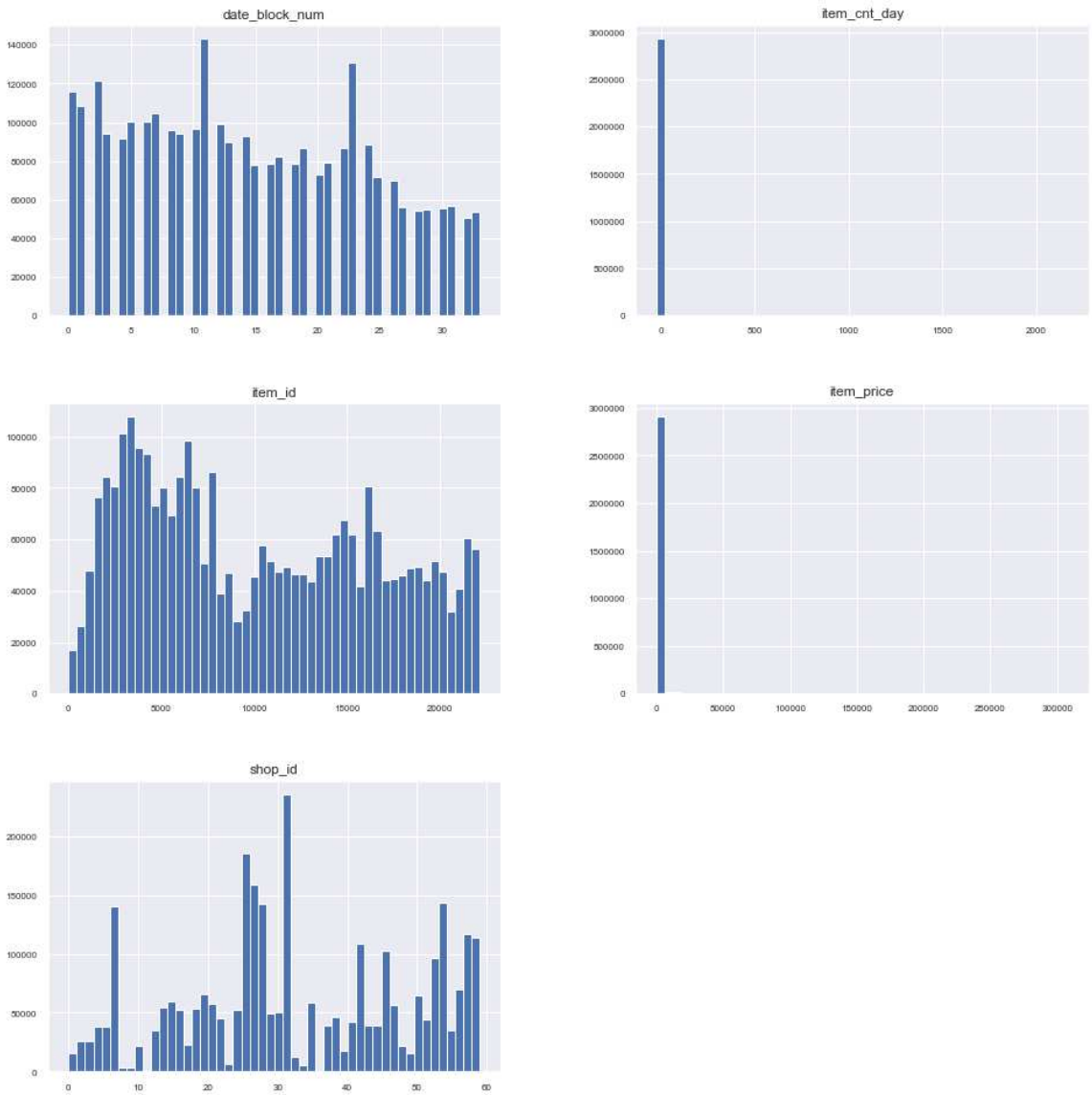


Data Visualization

- Problem Statement
- Exploratory Data Analysis
- Data Information
- Data Visualization
- Data Visualization**
- Data Visualization
- Data Visualization
- Feature Engineering
- Methods
- Forecast Results
- Conclusion

Exp

It seems that item_id and shop_id has a huge impact on sales and sales tend to decline with the date.



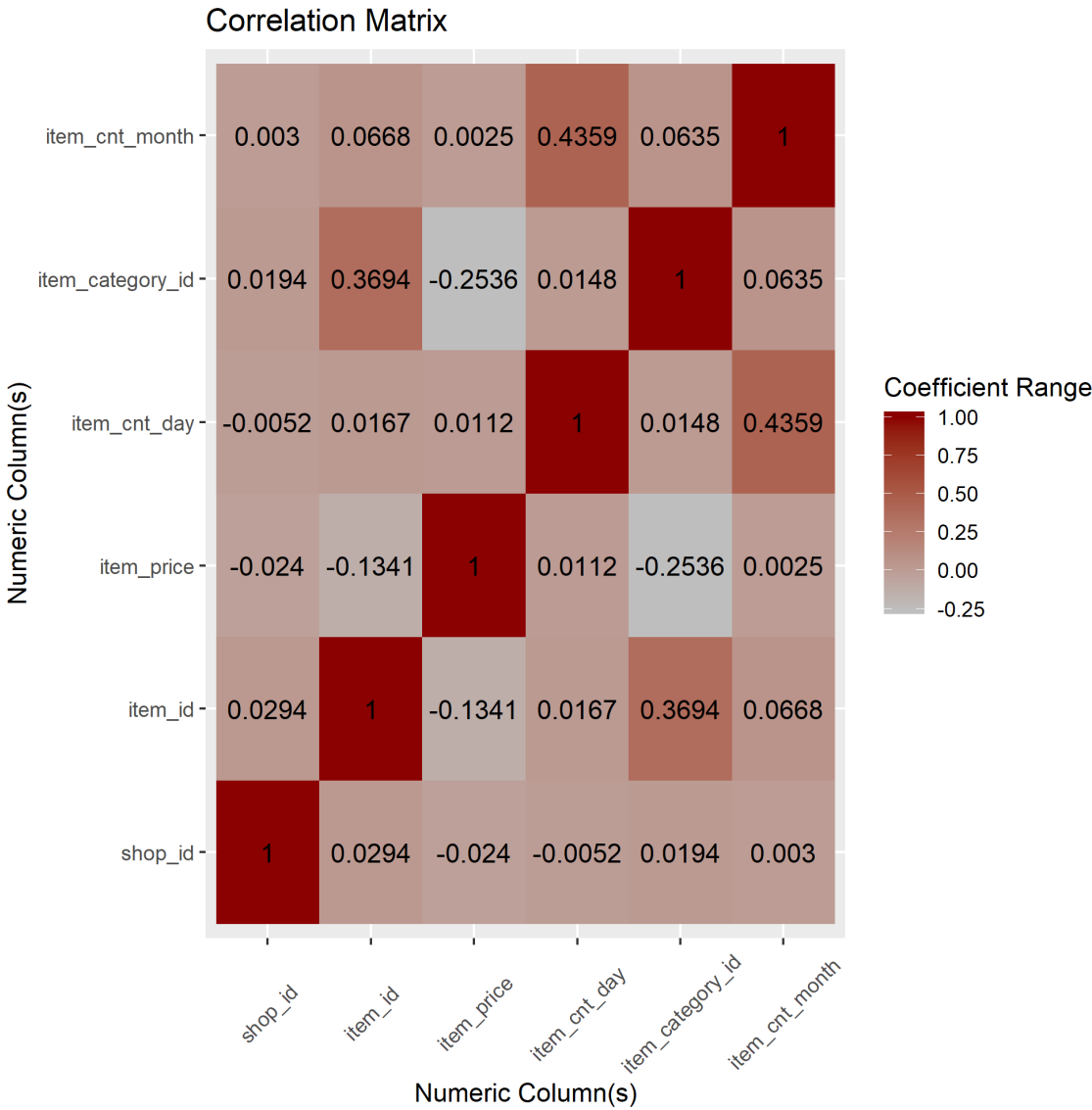
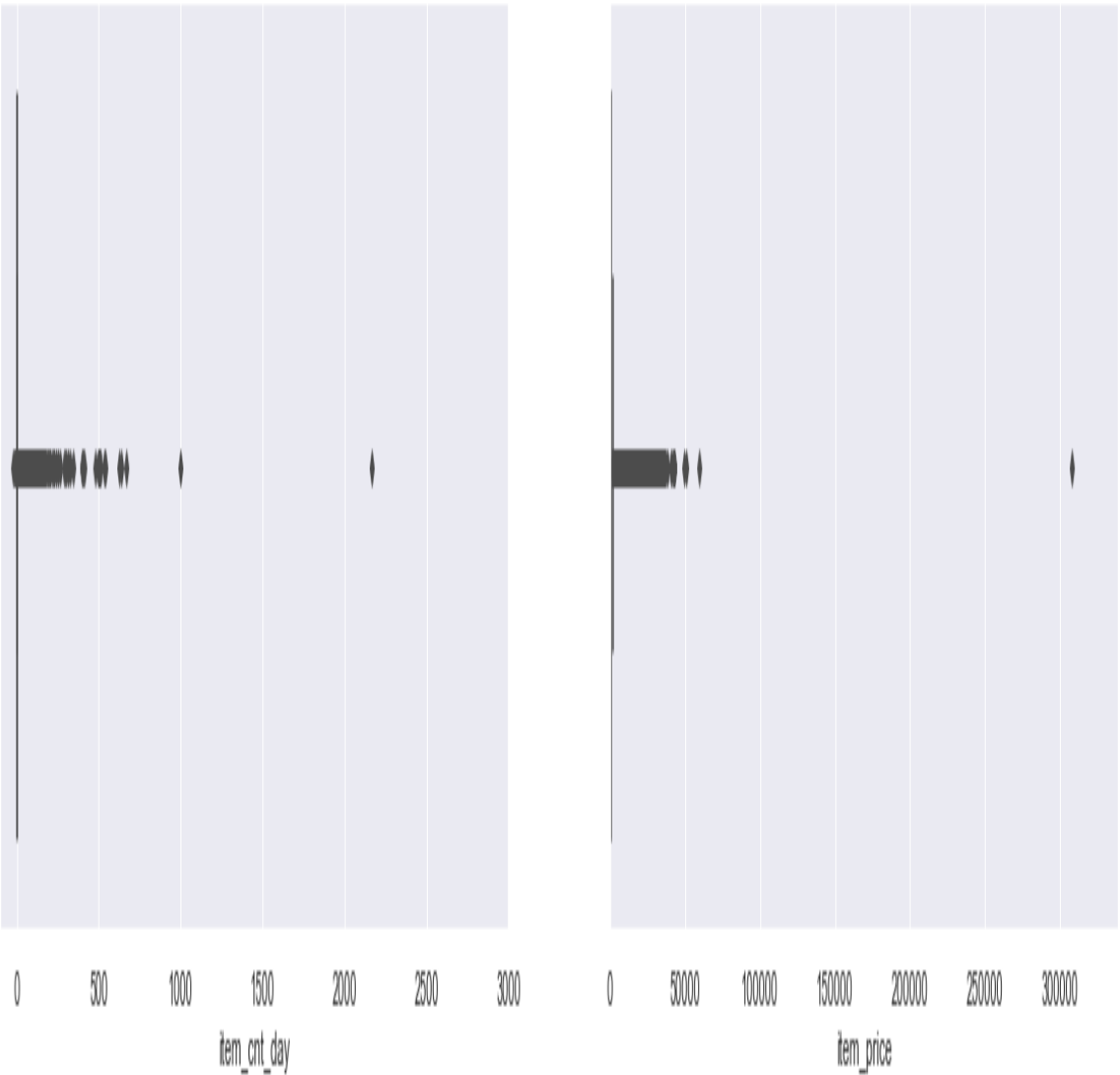


Data Visualization

- Problem Statement
- Exploratory Data Analysis
 - Data Information
 - Data Visualization
 - Data Visualization
 - Data Visualization**
 - Data Visualization
- Feature Engineering
- Methods
- Forecast Results
- Conclusion

Exp

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. We can see that the outliers are very small, so can be ignored.

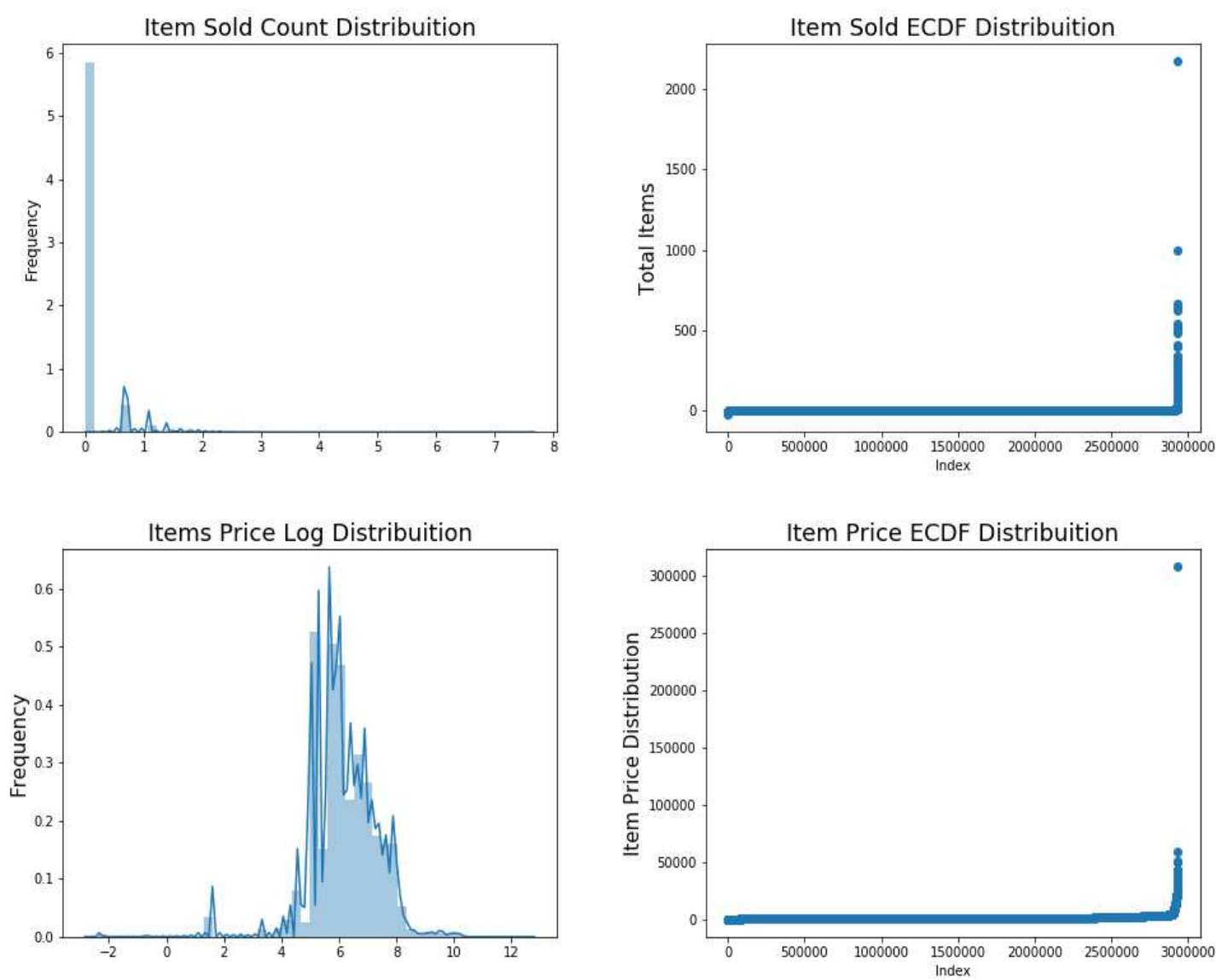




- Problem Statement
- Exploratory Data Analysis
- Data Information
- Data Visualization
- Data Visualization
- Data Visualization
- Data Visualization
- Feature Engineering
- Methods
- Forecast Results
- Conclusion

Exp

It can be seen from the scatter plot that the daily sales volume of the product is mainly concentrated between 0 and 1, and the price of the product can also be concentrated.





- [Problem Statement](#)
- [Exploratory Data Analysis](#)
- [Feature Engineering](#)**
- [Features importance](#)
- [Methods](#)
- [Forecast Results](#)
- [Conclusion](#)

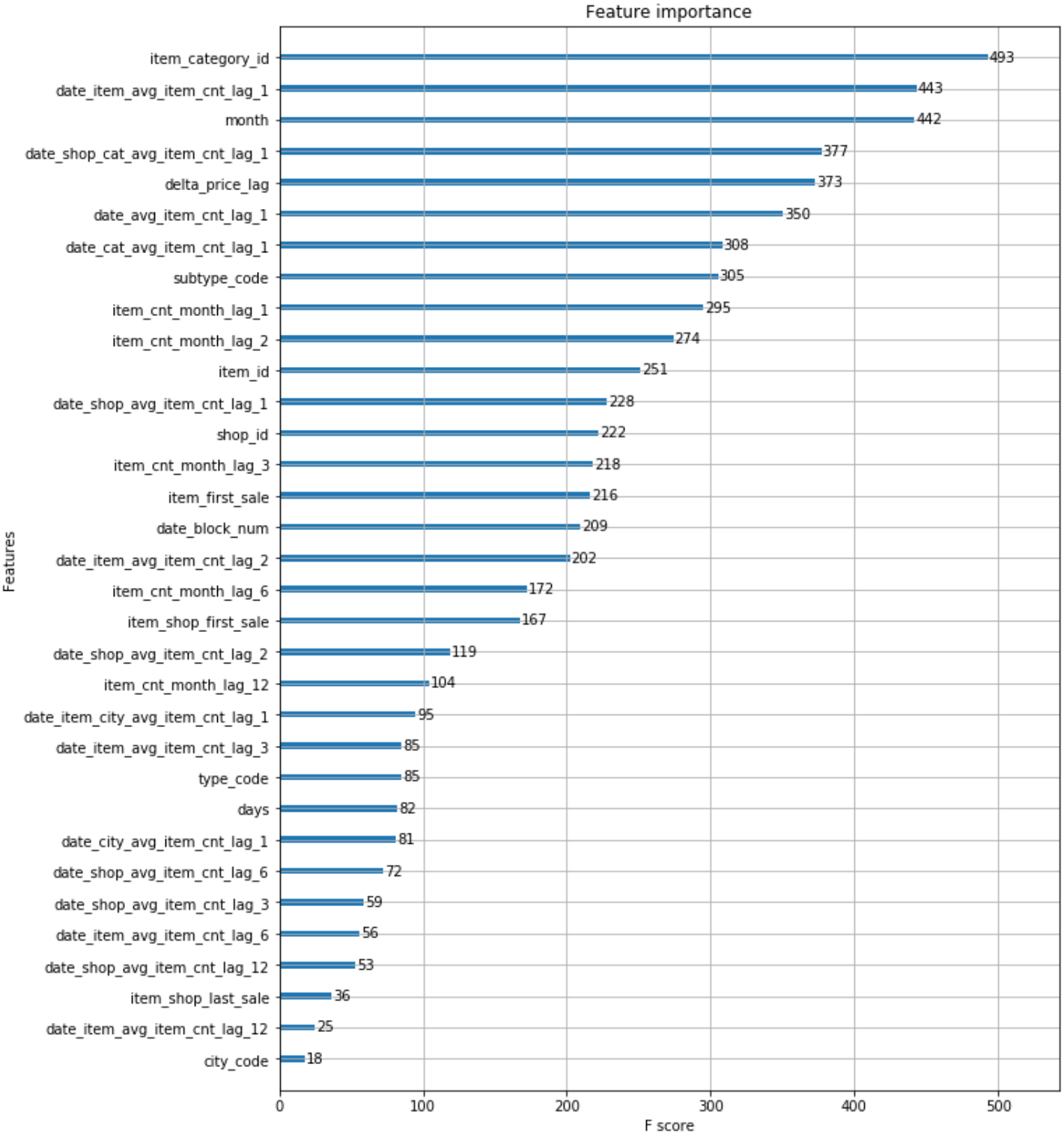
Feature Engineering



Features importance

- Problem Statement
- Exploratory Data Analysis
- Feature Engineering
- Features importance**
- Methods
- Forecast Results
- Conclusion

We take all of these features to form a new train datad.





- [Problem Statement](#)
- [Exploratory Data Analysis](#)
- [Feature Engineering](#)
- [Methods](#)**
- [Ensembling](#)
- [Route](#)
- [Forecast Results](#)
- [Conclusion](#)

Methods



Ensembling

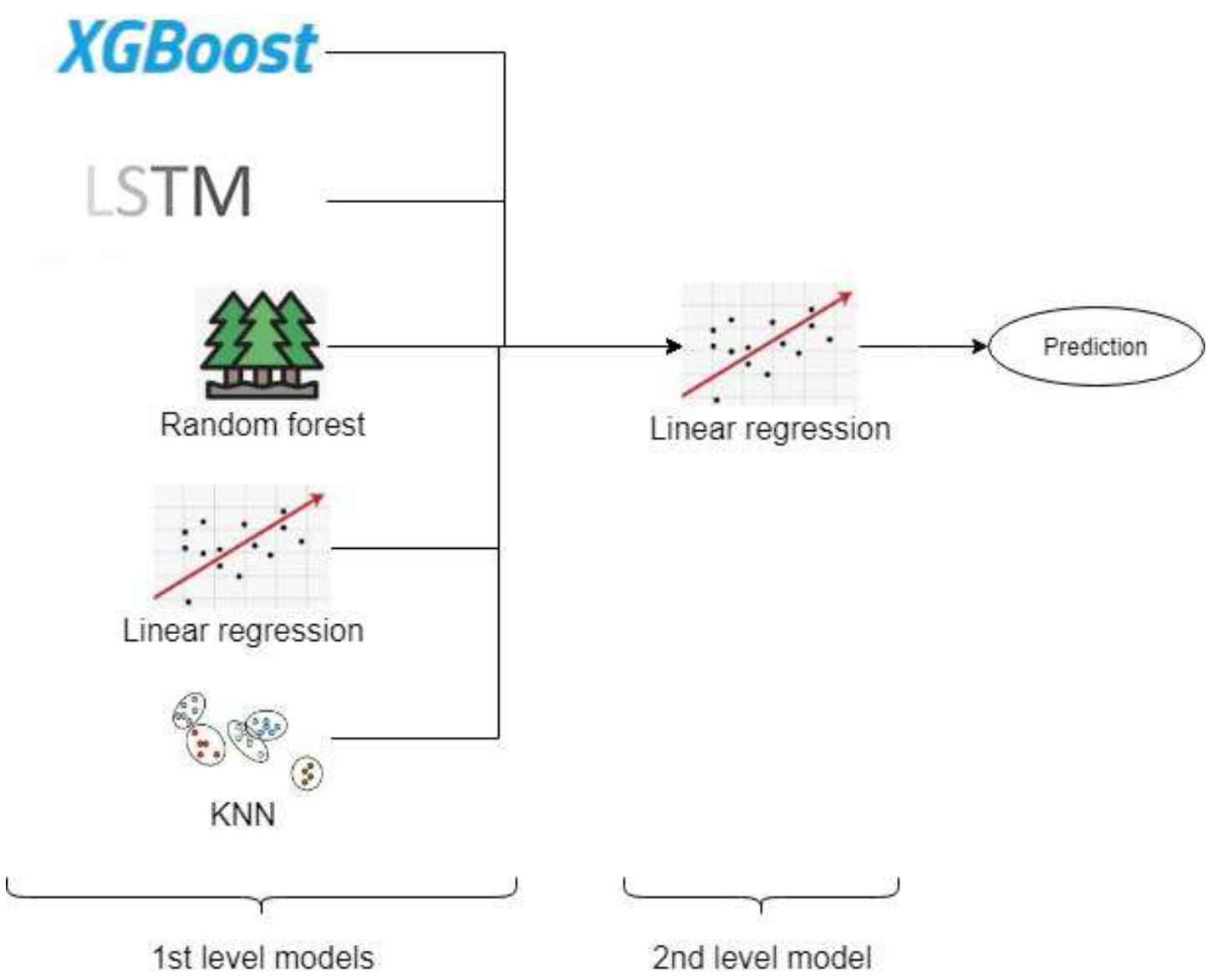
- [Problem Statement](#)
- [Exploratory Data Analysis](#)
- [Feature Engineering](#)
- [Methods](#)
- [Ensembling](#)**
- [Route](#)
- [Forecast Results](#)
- [Conclusion](#)

To combine the 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.

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

- Problem Statement
- Exploratory Data Analysis
- Feature Engineering
- Methods
- Ensembling
- Route
- Forecast Results
- Conclusion

Here is an image to help the understanding





[Problem Statement](#)

[Exploratory Data Analysis](#)

[Feature Engineering](#)

[Methods](#)

[Forecast Results](#)

[Evaluation Methods](#)

[Forecast Results](#)

[Conclusion](#)

Forecast Results



Evaluation Methods

- [Problem Statement](#)
- [Exploratory Data Analysis](#)
- [Feature Engineering](#)
- [Methods](#)
- [Forecast Results](#)
- [Evaluation Methods](#)
- [Forecast Results](#)
- [Conclusion](#)

■ RMSE





Forecast Results

- Problem Statement
- Exploratory Data Analysis
- Feature Engineering
- Methods
- Forecast Results
- Evaluation Methods
- Forecast Results**
- Conclusion

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

Table 1: 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

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



[Problem Statement](#)

[Exploratory Data Analysis](#)

[Feature Engineering](#)

[Methods](#)

[Forecast Results](#)

Conclusion

Conclusion



Conclusion

- [Problem Statement](#)
- [Exploratory Data Analysis](#)
- [Feature Engineering](#)
- [Methods](#)
- [Forecast Results](#)
- [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.



Thank you & Question

