Store Sales Forecasting for Favorita

MACHINE LEARNING PROJECT BY XINGJI JAX LI

Introduction

- An online competition from Kaggle
- Problem Overview: Sales forecasting for Favorita, a big company base in Ecuador that sells all kinds of daily commodities.
- Objective: Build up **a robust machine learning project** which includes exploratory data analysis (EDA), feature engineering, model building and comparing, and result conclusion.

Data:

Raw data files used:

Train.csv

Sales records:
Date, store,
Family(product kind)

oil.csv

Daily oil price:
Date, oil price

store.csv

Store information:
City, state,
store types

Holiday.csv

Holiday information:
Dates and names of
local holidays

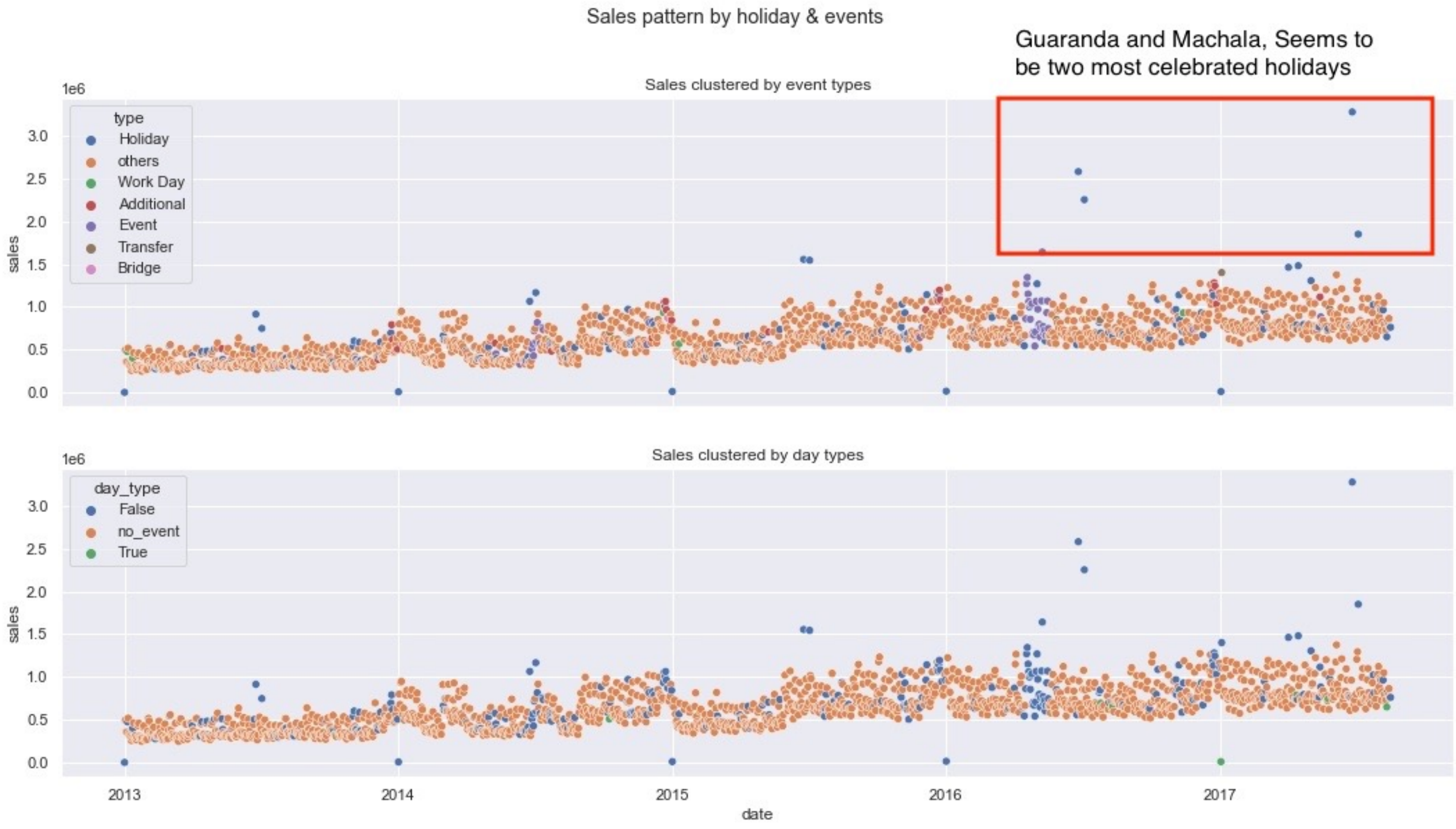
EDA: sales and oil price correlation

- Ecuador is an oil-dependent country and its economical health is highly vulnerable to shocks in oil prices
- Negative relation in yearly scope
- Alignment in big drops

EDA: sales clustered by store features

- Type A is out standing, while others are more blended
- The trends among different types of stores are similar
- It can be a good idea to add the store type as a feature
- It's possible to build an individual model for A
- More clusters were tried, but not shown here

EDA: sales in holidays



- Sales do not differ a lot among holidays and normal days
- But there are two outstanding holidays which lift the sales a lot.
- It can be a good idea to add a dummy feature for holiday which only indicates those two.

Methodology : feature engineering

	date	store_nbr	family	sales	onpromotion	city	state	type	cluster	oil_price	holiday
0	2013-01-01	1	AUTOMOTIVE	0.0	0	Quito	Pichincha	D	13	93.14	0.0
1	2013-01-01	1	BABY CARE	0.0	0	Quito	Pichincha	D	13	93.14	0.0
2	2013-01-01	1	BEAUTY	0.0	0	Quito	Pichincha	D	13	93.14	0.0
3	2013-01-01	1	BEVERAGES	0.0	0	Quito	Pichincha	D	13	93.14	0.0
4	2013-01-01	1	BOOKS	0.0	0	Quito	Pichincha	D	13	93.14	0.0

Categorical

One Hot Encoding

	date	store_nbr	sales	onpromotion	oil_price	holiday	type_A	type_B	type_C	type_D	...	cluster_8
0	2013-01-01	1	0.0	0	93.14	0.0	0	0	0	1	...	0
1	2013-01-01	1	0.0	0	93.14	0.0	0	0	0	1	...	0
2	2013-01-01	1	0.0	0	93.14	0.0	0	0	0	1	...	0
3	2013-01-01	1	0.0	0	93.14	0.0	0	0	0	1	...	0

One hot encoding:
type = A / B / C / D

type_A = 0 / 1
type_B = 0 / 1
type_C = 0 / 1
type_D = 0 / 1

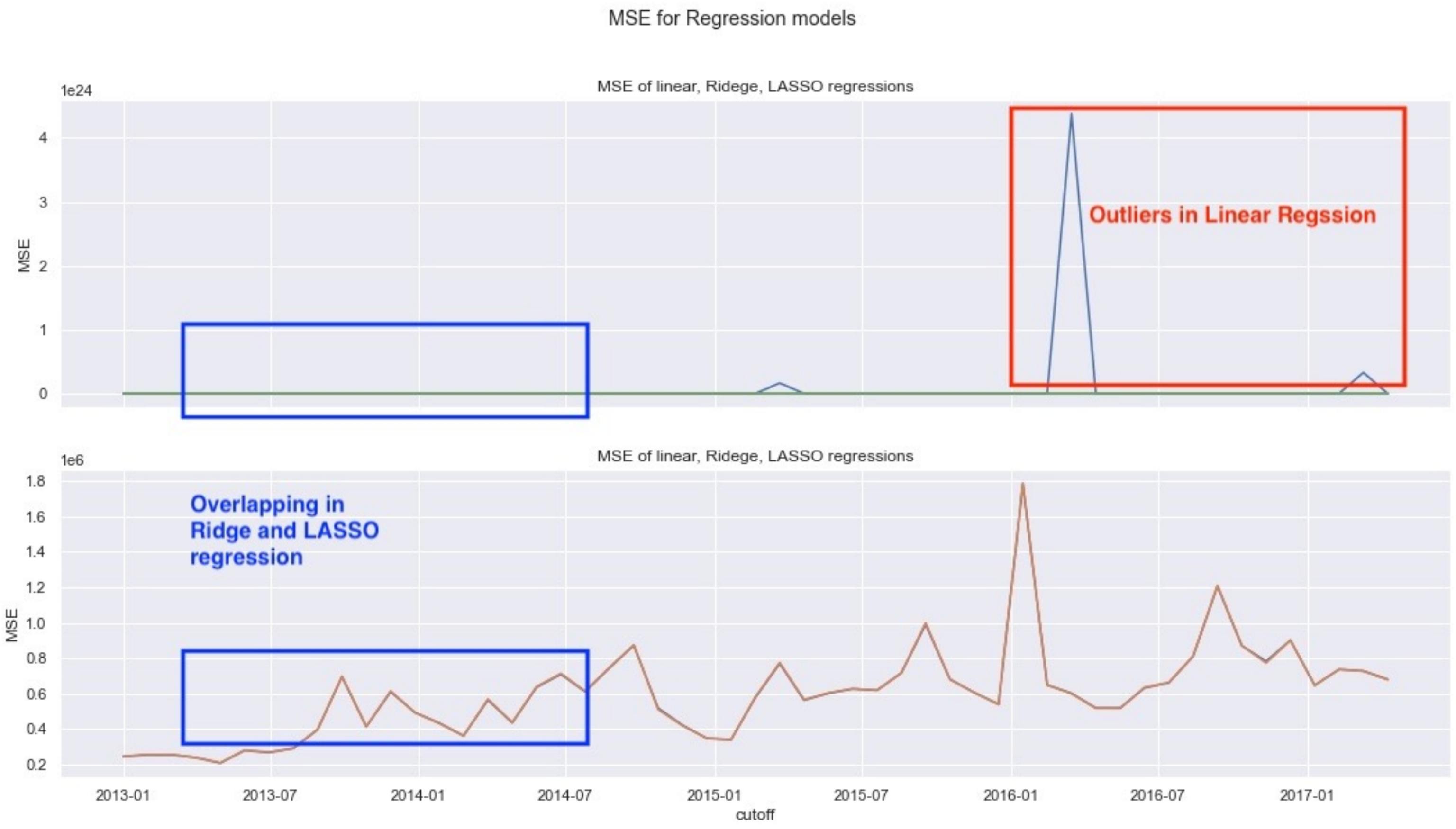
Methodology : cutoffs

Simple Idea:

Applying Cutoffs: check the robustness of model performance

Default Stride = Training range

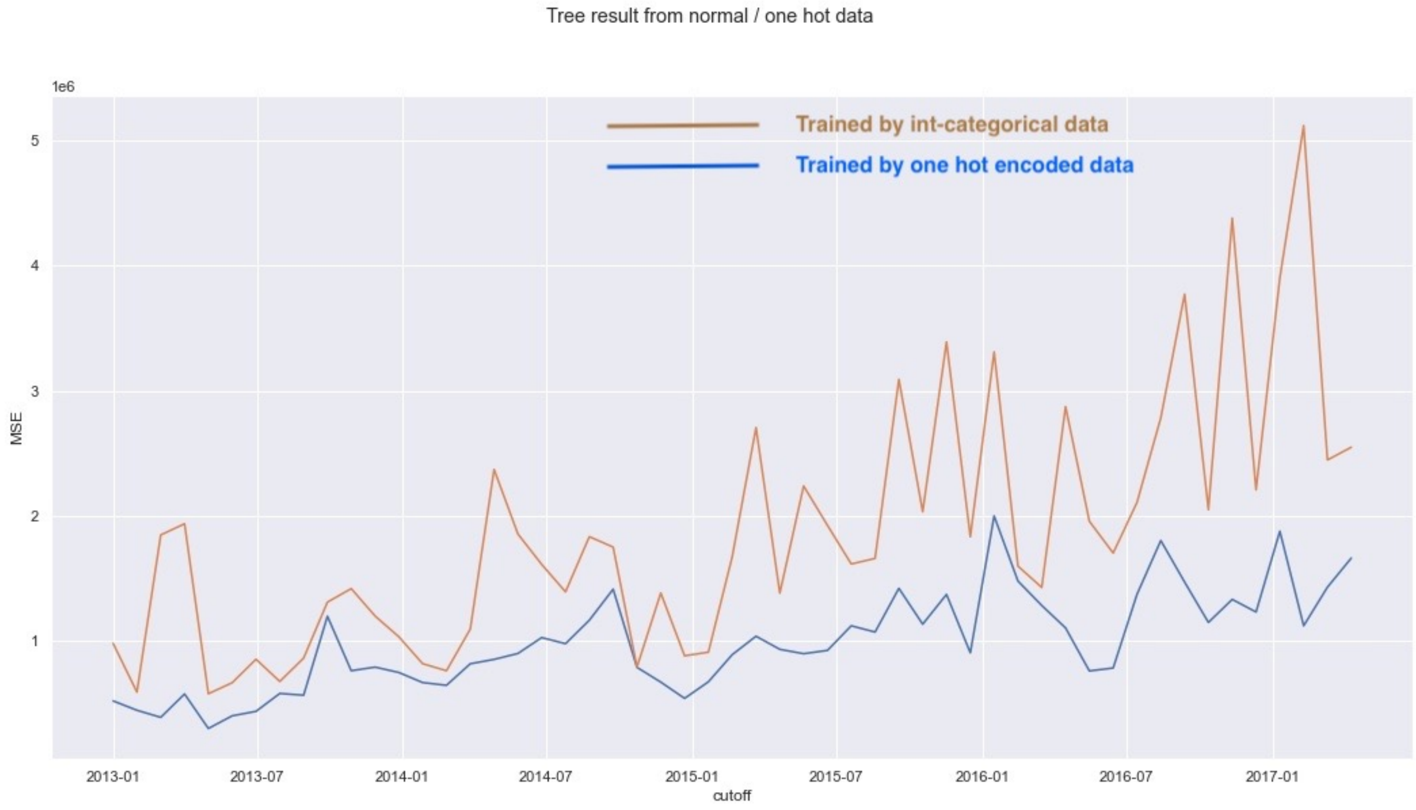
Methodology : regressions



- Massive outliers in simple linear regression model
- High overlapping among Ridge and LASSO regressions

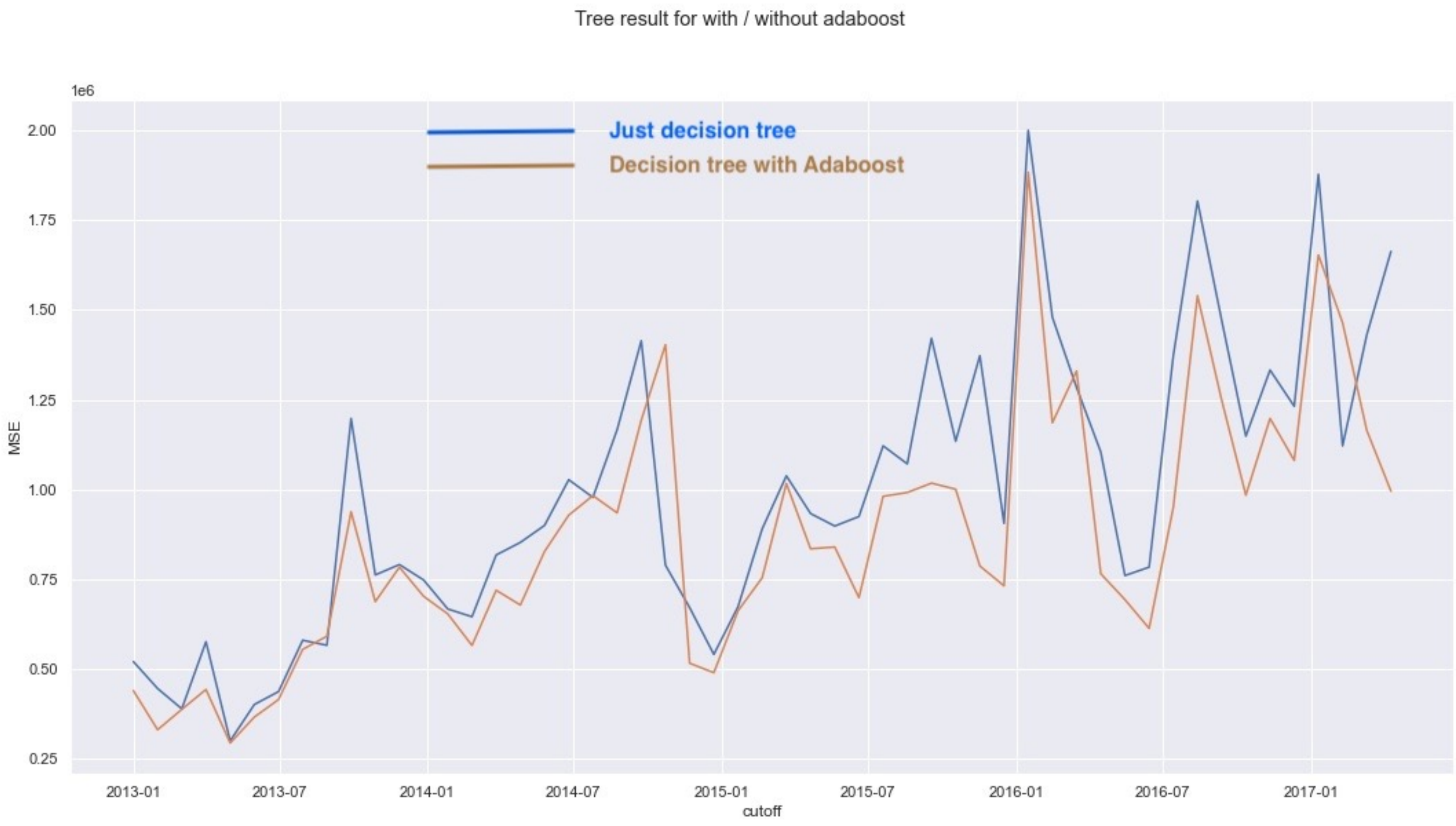
Methodology : decision trees

Comparing data used:



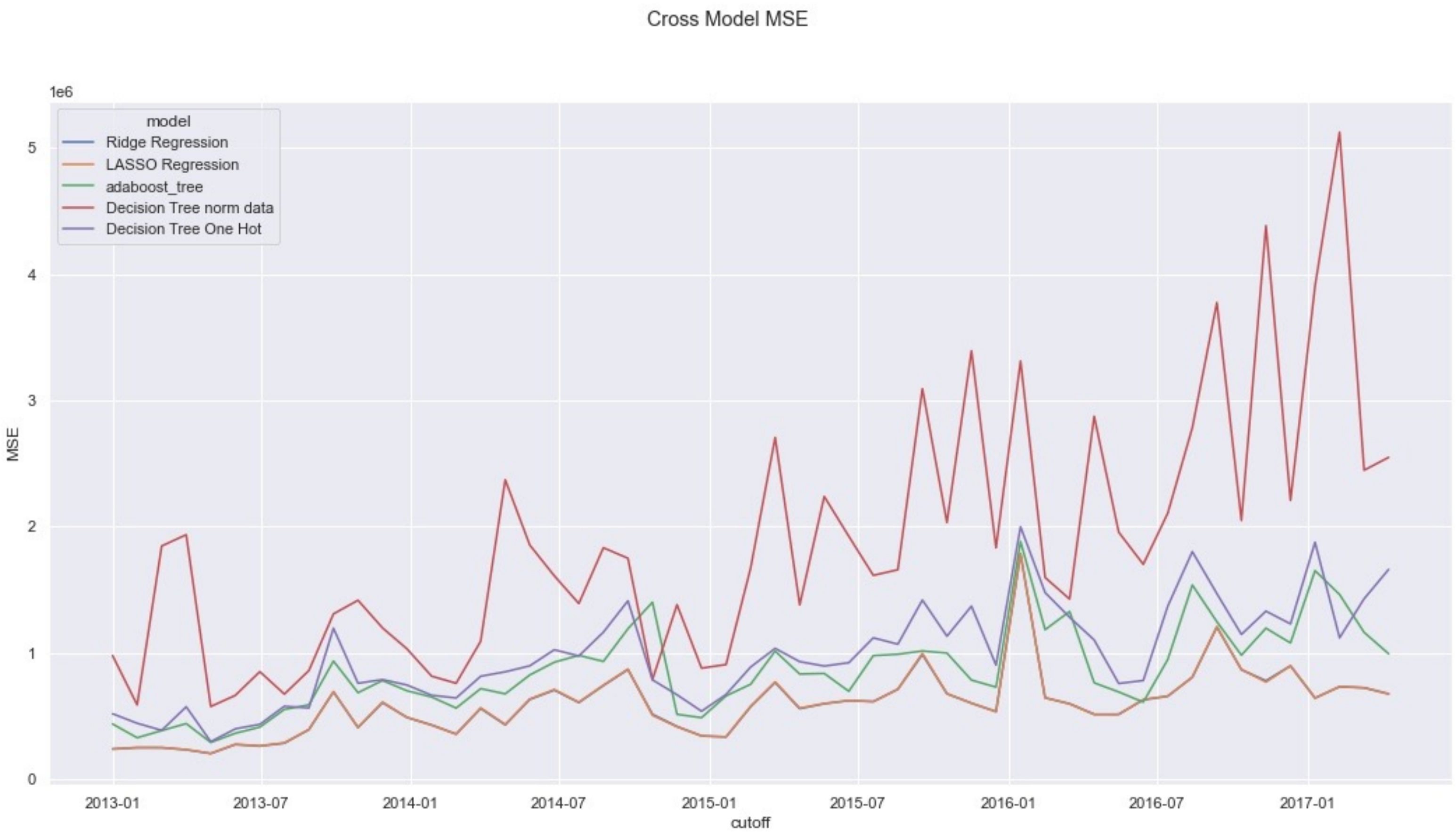
- Tree structure is able to take integer encoded categorical data
- The overall performance is worse than using one hot encoded data

Methodology : decision trees - Adaboost



- Adaboost can improve the result slightly.
- But it take much longer to run.

Conclusion : cross model evaluation



- Generally, it is better to use one hot encoded training data
- Despite the decision tree models are a bit more complex, the ridge and LASSO regression would still reach to a slightly better result.
- Either LASSO or ridge regression can be a preferable choice in application

Conclusion : next steps

- RNN (on going)
- Other improvement methods:
  - Tuning hyperparameters (cut off time ranges, max depths in tree models …)
  - Other ensemble methods, for example, build another model to weighted combine the result we get from different models.