FINAL PROJECT

INFO6105

KAGGLE: Walmart Recruiting

Store Sales Forecasting

Bodong Zhou
Yisheng Yang
Qing Hu

Dec 9, 2019

CONTENT

PART ONE
INTRODUCTION

PART TWO
DATA SUMMARY

PART THREE
METHODOLOGY

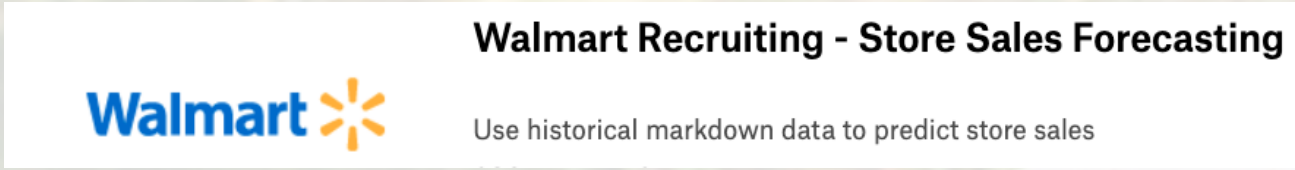
PART FOUR
CONSEQUENCE

01

PART ONE

INTRODUCTION

INTRODUCTION



The problem is to predict weekly sales data based on historical sales data for 45 Walmart stores located in different regions.

If we are able to predict the weekly sales we can use this knowledge to better manage the supply chain and we can also see how each departments are affected by the markdown and the extent of the impact.

02

PART TWO

DATA SUMMARY

S U M M A R Y

The data contains 8191 rows with 12 features each row.

Major features

- Store - the store number
- Date - the week
- Temperature - average temperature in the region
- Fuel_Price - cost of fuel in the region
- MarkDown1-5 - anonymized data related to promotional markdowns that Walmart is running. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA.
- CPI - the consumer price index
- Unemployment - the unemployment rate
- IsHoliday - whether the week is a special holiday week

S U M M A R Y

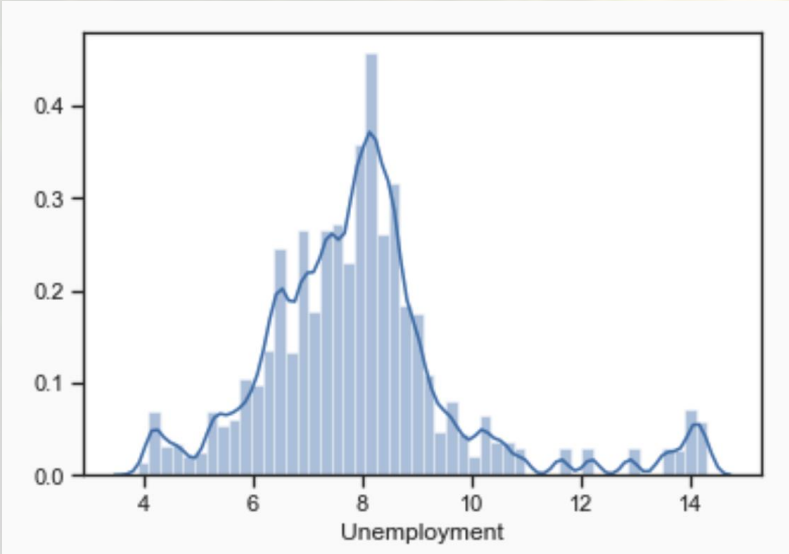
In addition, Walmart runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of which are the Super Bowl, Labor Day, Thanksgiving, and Christmas. The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks. Part of the challenge presented by this competition is modeling the effects of markdowns on these holiday weeks in the absence of complete/ideal historical data.

DATA ANALYSIS

We perform 4 common techniques and found several patterns

Distribution of the dataset

DATA ANALYSIS



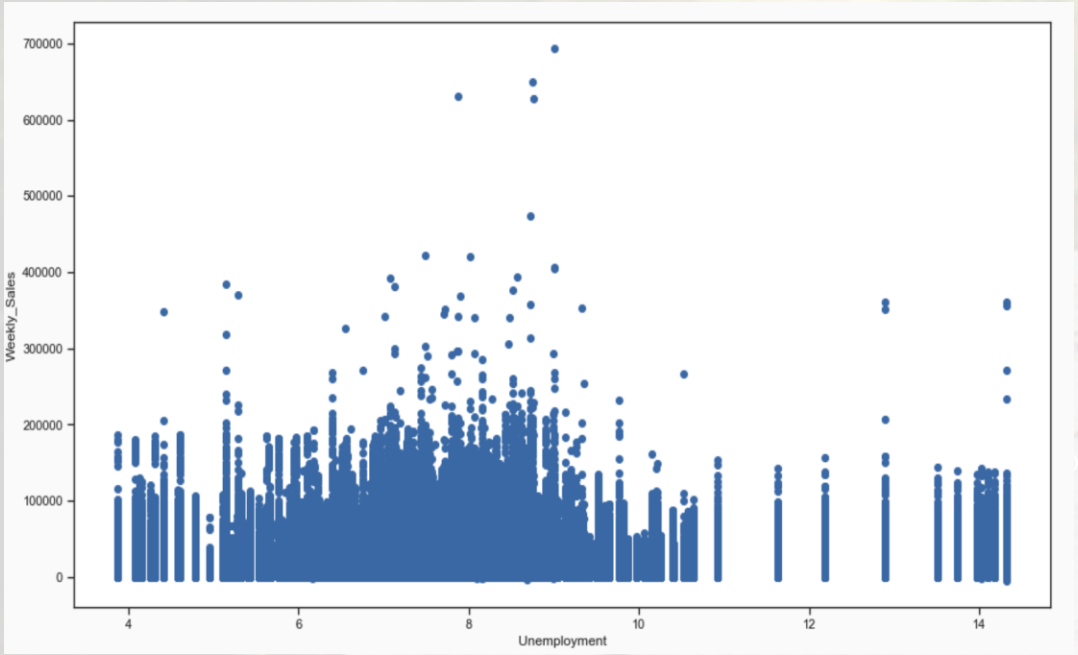
We found the unemployment rate is also not balanced; the peak is at 8%
Also, the type of stores are unbalanced. there are more type A store than B and C combined

DATA ANALYSIS

We perform 4 common techniques and found several patterns

Relations between different features and weekly sales

DATA ANALYSIS



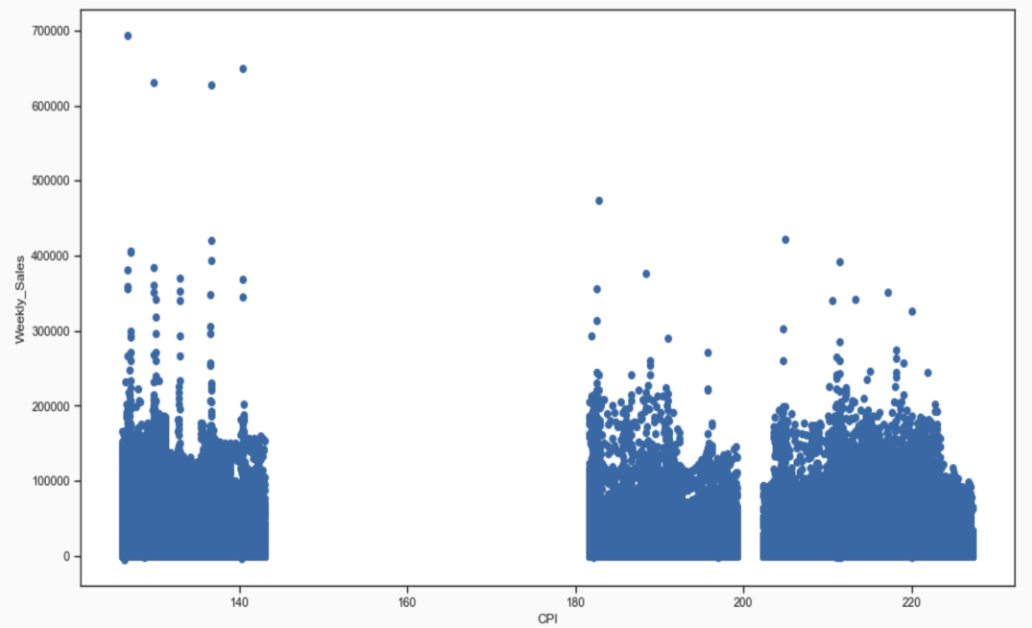
Although there are so many data points, we can still see that high sales are closely related to low unemployment rates.

DATA ANALYSIS

We perform 4 common techniques and found several patterns

Relations between CPIs and sales

DATA ANALYSIS



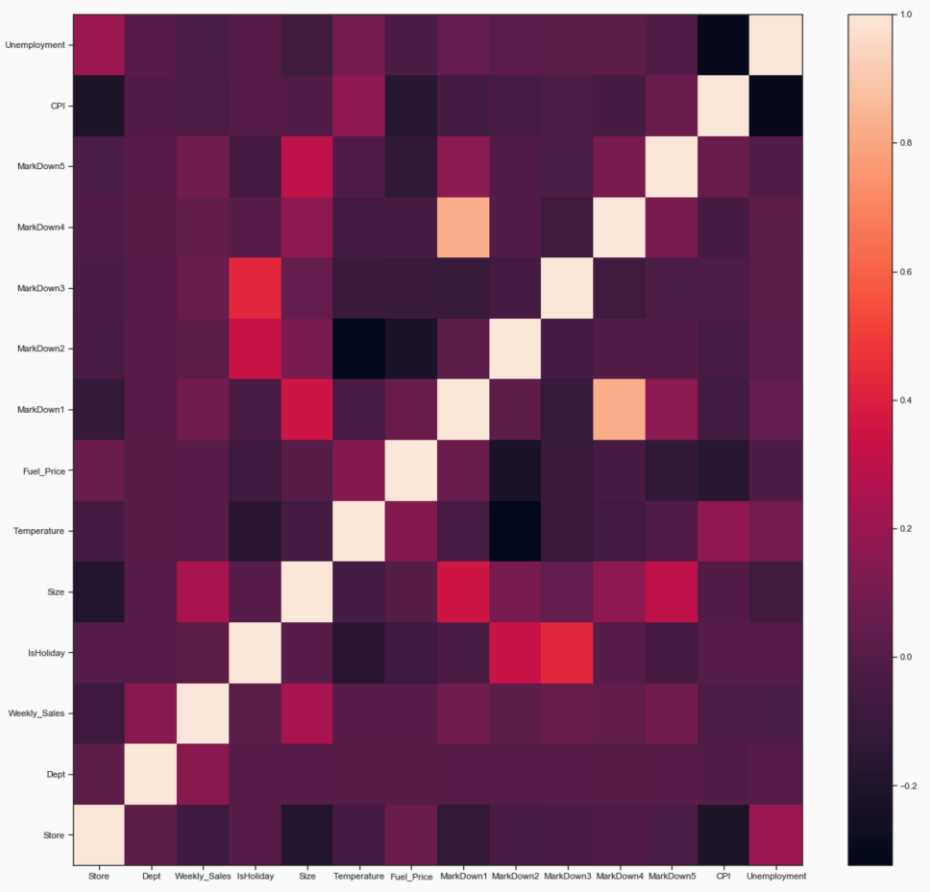
we noticed that lower CPIs creates more sales. And there seems to be no data in the CPI range of [143, 181]

DATA ANALYSIS

We perform 4 common techniques and found several patterns

Correlation **coefficient between** different **features**

DATA ANALYSIS



We focus on the correlation between weekly sales and other factors. It appears that the size of the store is very relevant to the weekly sales.

(*We only include significant findings in this paragraph, other explorations can be found in the notebook* data-analysis.ipynb)

03

PART THREE

METHODOLOGY

METHODOLOGY

Reading the data

Reading the data from train.csv, features.csv, stores.csv, test.csv.

METHODOLOGY

Preprocessing

Drop “IsHoliday” feature from features.csv. Merge train.csv, features.csv and stores.csv as train dataset. Merge test.csv, features.csv and stores.csv as test dataset.

There are many null values for the field Markdown(Because markdowns are special events that are not held every day, we replace these values with 0)

	Store	Dept	weeklySales	isHoliday	Size	Temperature	MarkDown1	MarkDown2	MarkDown4	MarkDown5	Month
0	1	1	24924.50	False	0	42.31	0.0	0.0	0.0	0.0	2
1	1	1	46039.49	True	0	38.51	0.0	0.0	0.0	0.0	2
2	1	1	41595.55	False	0	39.93	0.0	0.0	0.0	0.0	2
3	1	1	19403.54	False	0	46.63	0.0	0.0	0.0	0.0	2
4	1	1	21827.90	False	0	46.50	0.0	0.0	0.0	0.0	3

METHODOLOGY

Train-valid split

We split train dataset into small size that containing 10,000 rows in order to find best estimator fast. Otherwise, it will take whole day to optimize hyper parameters. Then, we split train dataset into valid dataset and another train dataset. The valid dataset may contain about 100,000 rows. This is for checking metrics to judge the performance of our three methods

METHODOLOGY

Preprocessing

With smaller size train dataset, we use GridSearchCV function to search best hyper parameters. During this process, we use 5 times cross validation to improve the precision. And we use mean absolute error metrics to judge whether we find the best one.

METHODOLOGY

THREE MODELS

EXTRA TREES

KNN

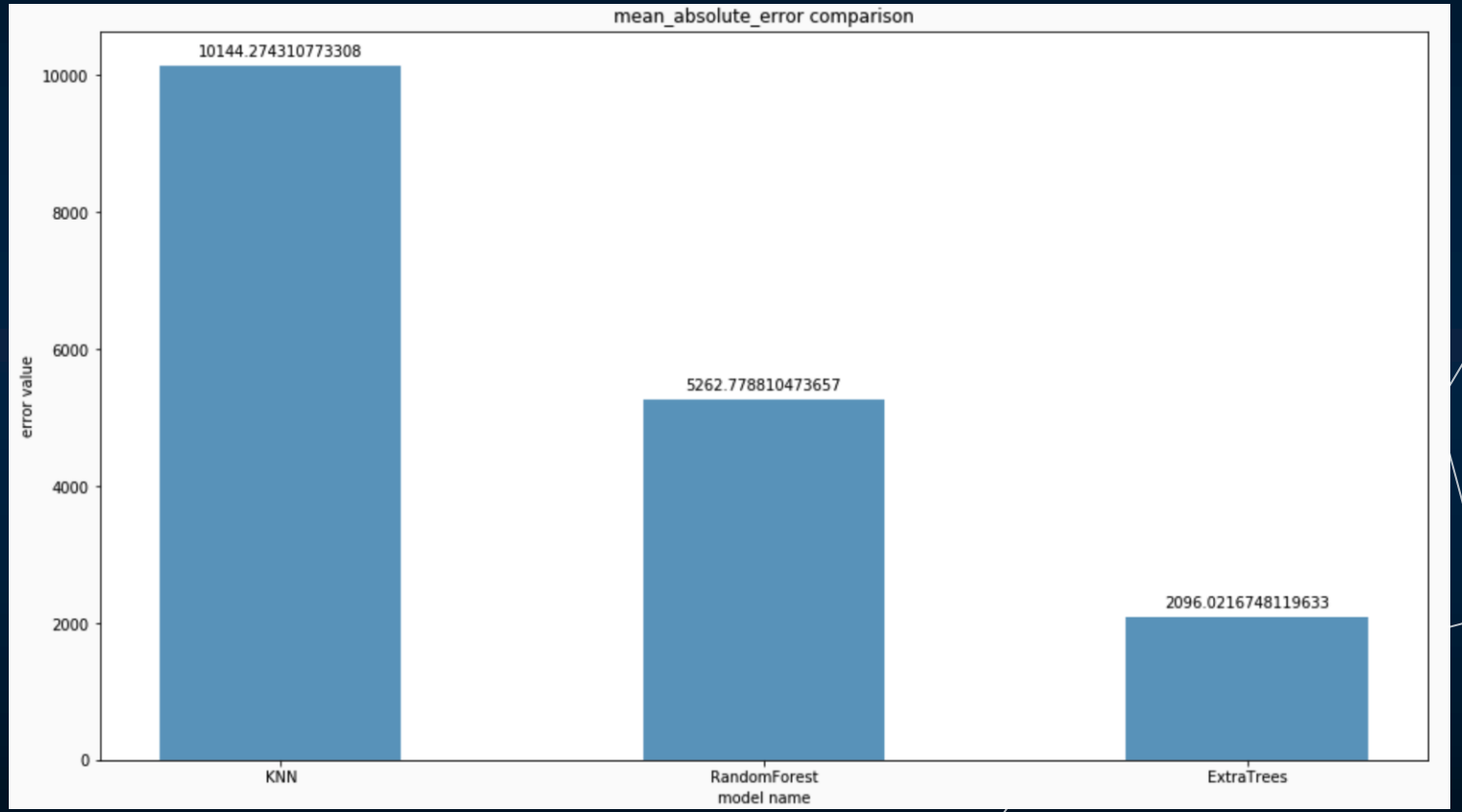
RANDON FOREST


04
PART FOUR
CONSEQUENCE

T W O M E T R I C S

MEAN ABSOLUTE ERROR

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

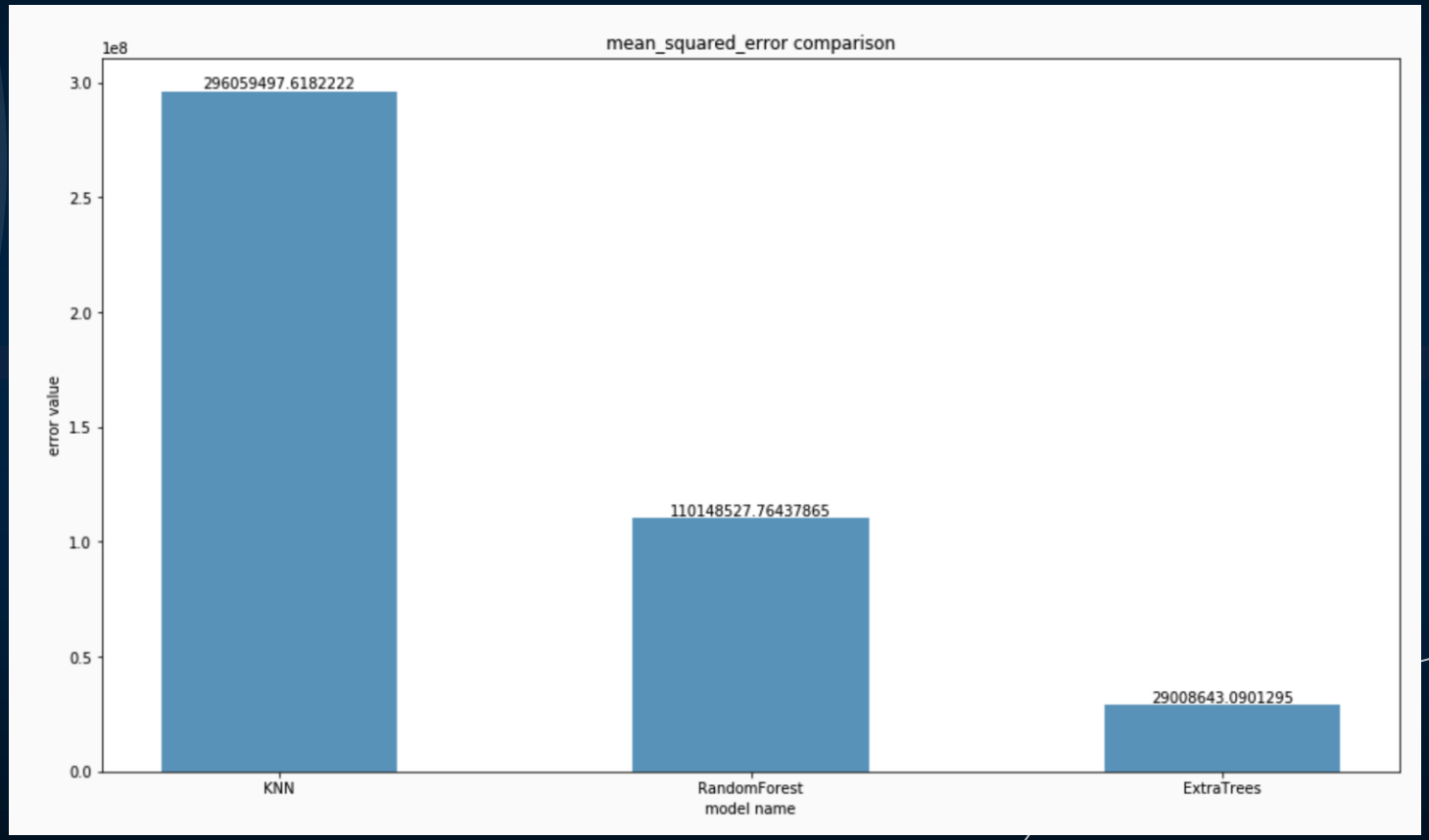


T W O M E T R I C S

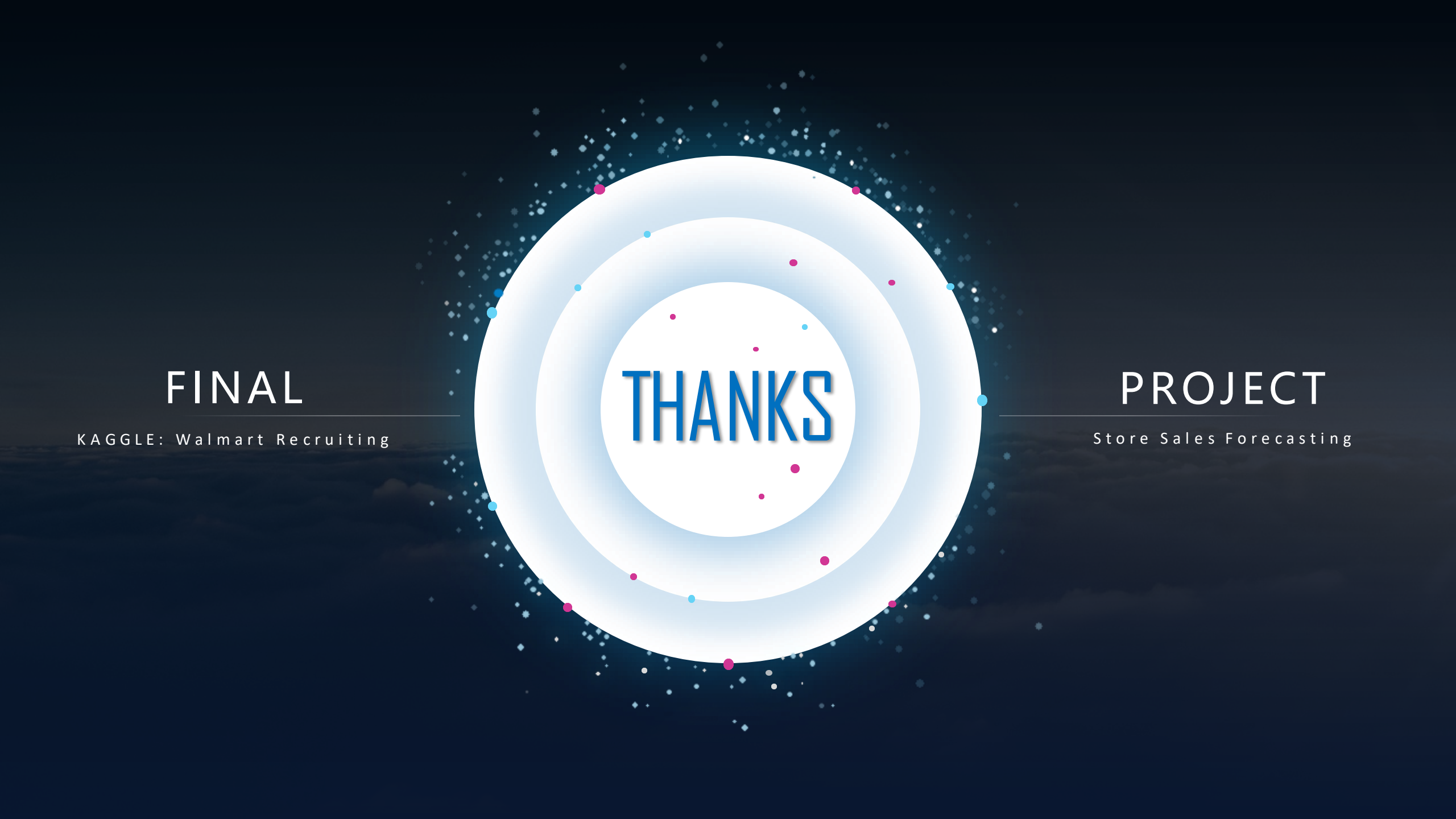
• • •

MEAN SQUARED ERROR

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$



(*We only include significant findings in this paragraph, other explorations can be found in the notebook* model-training.ipynb)


FINAL
KAGGLE: Walmart Recruiting
THANKS
PROJECT
Store Sales Forecasting