

Project Overview

Liyu Chen

Overview

- Logistic
- ML Advice

Store Sales - Time Series Forecasting

In this “getting started” competition, you’ll use time-series forecasting to forecast store sales on data from Corporación Favorita, a large Ecuadorian-based grocery retailer.

Specifically, you'll build a model that more accurately predicts the unit sales for thousands of items sold at different Favorita stores. You'll practice your machine learning skills with an approachable training dataset of dates, store, and item information, promotions, and unit sales.

Teams

- The project should be done in teams comprising of 3-4 members (preferably 4 members).
- Fill out your team information in the provided google form and create Kaggle accounts by **11:59 pm, November 11, 2022**. Make sure that all members of your team are registered under a single team name which begins with CSCI567_id[TEAM ID], for example, CSCI567_id16.
- Team can have members from different sections of the class (offline, online, DEN).

Grading

- Relative rank on the leaderboard (40%) + the project report and code (60%).
- Only take the relative ranking among the class' teams into consideration
- Members of the same team will receive the same scores
- Bonus points: you earn 10 pts if you are
 - the first 5 teams among all teams in CSCI-567 class at the HW4 submission deadline (Nov 16)
 - the first 5 teams among all teams in CSCI-567 class in the final leaderboard (Dec 11)
 - the team that wins the 1st place among all teams in CSCI-567 class
 - the team wins the 1st place among all the teams on the leaderboard

Bonus points above are cumulative.

Deliverables

- Each team needs to write the project report in NeurIPS format. (6 pages maximum, including references; this page limit is strict)
- In your report, you should cover the details of your solutions.
- Use Python as the programming language. You are allowed to use public available computational resources.

Policy on collaboration

- In line with the rules of the competition, you are only allowed to share code within your own team.
- Discussion about approaches between each team members and cross-teams are allowed and we encourage you to actively engage in forums, piazza, and discussion with the Kaggle's community

Useful References on Kaggle

- Pandas tutorial: <https://www.kaggle.com/learn/pandas>
- Time series course: <https://www.kaggle.com/learn/time-series>

General ML Advice: 7 Steps of ML Systems

Step 1: Acquire Data

Step 2: Look at your data* -- after every step.

Step 3: Create train/dev/test splits

Step 4: Create/Refine a specification

Step 5: Build model (simplest that works!)

Step 6: Measurement

Step 7: Repeat.

Source: <https://drive.google.com/file/d/1zEPVIBTtpFJi-y0lwjab4-AFcjkEdMOu/view?usp=sharing>

General ML Advice: 7 Steps of ML Systems

Step 1: Acquire Data

Step 2: Look at your data* -- after every step.

Step 3: Create train/dev/test splits

Step 4: Create/Refine a specification

Step 5: Build model (simplest that works!)

Step 6: Measurement

Step 7: Repeat.

Source: <https://drive.google.com/file/d/1zEPVIBTtpFJi-y0lwjab4-AFcjkEdMOu/view?usp=sharing>

Look at your data

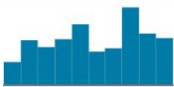

- Very important! The preliminary of feature engineering
- You should do this after every iteration, which could give you inspiration in improving your results
- Let's look at the data together

holidays_events.csv (22.31 kB)



Detail Compact Column

6 of 6 columns ▾

date	type	locale	locale_name	description	transferred
	Holiday 63% Event 16% Other (73) 21%	National 50% Local 43% Other (24) 7%	Ecuador 50% Quito 4% Other (163) 47%	Carnaval 3% Fundacion de Cuenca 2% Other (333) 95%	 <div> true 12 3% false 338 97% </div>
2012-03-02	Holiday	Local	Manta	Fundacion de Manta	False
2012-04-01	Holiday	Regional	Cotopaxi	Provincializacion de Cotopaxi	False
2012-04-12	Holiday	Local	Cuenca	Fundacion de Cuenca	False
2012-04-14	Holiday	Local	Libertad	Cantonizacion de Libertad	False
2012-04-21	Holiday	Local	Riobamba	Cantonizacion de Riobamba	False
2012-05-12	Holiday	Local	Puyo	Cantonizacion del Puyo	False
2012-06-23	Holiday	Local	Guaranda	Cantonizacion de Guaranda	False
2012-06-25	Holiday	Regional	Imbabura	Provincializacion de Imbabura	False
2012-06-25	Holiday	Local	Latacunga	Cantonizacion de	False

Data Explorer

124.76 MB



- holidays_events.csv
- oil.csv
- sample_submission.csv
- stores.csv
- test.csv
- train.csv
- transactions.csv

Summary

- 7 files
- 29 columns

oil.csv (20.58 kB)

Detail Compact Column

date	# dcoiltwico
	
2013-01-01	
2013-01-02	93.14
2013-01-03	92.97
2013-01-04	93.12
2013-01-07	93.2
2013-01-08	93.21
2013-01-09	93.08
2013-01-10	93.81
2013-01-11	93.6
2013-01-14	94.27
2013-01-15	93.26
2013-01-16	94.28

Download Split View Expand

2 of 2 columns

Data Explorer

124.76 MB

- holidays_events.csv
- oil.csv
- sample_submission.csv
- stores.csv
- test.csv
- train.csv
- transactions.csv

Summary

- 7 files
- 29 columns

sample_submission.csv (342.15 kB)



Detail Compact Column

id	# sales
3000888	0.0
3000889	0.0
3000890	0.0
3000891	0.0
3000892	0.0
3000893	0.0
3000894	0.0
3000895	0.0
3000896	0.0
3000897	0.0
3000898	0.0
3000899	0.0

2 of 2 columns

Data Explorer

124.76 MB

- holidays_events.csv
- oil.csv
- sample_submission.csv
- stores.csv
- test.csv
- train.csv
- transactions.csv

Summary

- 7 files
- 29 columns

stores.csv (1.39 kB)

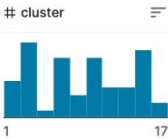


Detail Compact Column

5 of 5 columns ▾



# store_nbr	city	state	type
1	Quito	Pichincha	D
2	Quito	Pichincha	D
3	Quito	Pichincha	D
4	Quito	Pichincha	D
5	Santo Domingo	Santo Domingo de los Tsachilas	D
6	Quito	Pichincha	D
7	Quito	Pichincha	D
8	Quito	Pichincha	D
9	Quito	Pichincha	B
10	Quito	Pichincha	C
11	Cayambe	Pichincha	B



# store_nbr	city	state	type	# cluster
1	Quito	Pichincha	D	13
2	Quito	Pichincha	D	13
3	Quito	Pichincha	D	8
4	Quito	Pichincha	D	9
5	Santo Domingo	Santo Domingo de los Tsachilas	D	4
6	Quito	Pichincha	D	13
7	Quito	Pichincha	D	8
8	Quito	Pichincha	D	8
9	Quito	Pichincha	B	6
10	Quito	Pichincha	C	15
11	Cayambe	Pichincha	B	6

Data Explorer

124.76 MB

- holidays_events.csv
- oil.csv
- sample_submission.csv
- stores.csv
- test.csv
- train.csv
- transactions.csv

Summary

- 7 files
- 29 columns

test.csv (1.02 MB)



Detail Compact Column

5 of 5 columns

id	date	# store_nbr	family	# onpromotion
			<div>33 unique values</div>	
3000888	2017-08-16	1	AUTOMOTIVE	0
3000889	2017-08-16	1	BABY CARE	0
3000890	2017-08-16	1	BEAUTY	2
3000891	2017-08-16	1	BEVERAGES	20
3000892	2017-08-16	1	BOOKS	0
3000893	2017-08-16	1	BREAD/BAKERY	12
3000894	2017-08-16	1	CELEBRATION	0
3000895	2017-08-16	1	CLEANING	25
3000896	2017-08-16	1	DAIRY	45
3000897	2017-08-16	1	DELI	18
3000898	2017-08-16	1	EGGS	1
3000899	2017-08-16	1	FROZEN FOODS	1

Data Explorer

124.76 MB

- holidays_events.csv
- oil.csv
- sample_submission.csv
- stores.csv
- test.csv
- train.csv
- transactions.csv

Summary

- 7 files
- 29 columns

train.csv (121.8 MB)



Detail Compact Column

6 of 6 columns

id	date	store_nbr	family	sales	onpromotion
			33 unique values		
0	31Dec12	1	AUTOMOTIVE	0.0	0
1	14Aug17	1	BABY CARE	0.0	0
2		1	BEAUTY	0.0	0
3		1	BEVERAGES	0.0	0
4		1	BOOKS	0.0	0
5		1	BREAD/BAKERY	0.0	0
6		1	CELEBRATION	0.0	0
7		1	CLEANING	0.0	0
8		1	DAIRY	0.0	0
9		1	DELI	0.0	0
10		1	EGGS	0.0	0
11		1	FROZEN FOODS	0.0	0

Data Explorer

124.76 MB

- holidays_events.csv
- oil.csv
- sample_submission.csv
- stores.csv
- test.csv
- train.csv
- transactions.csv

Summary

- 7 files
- 29 columns

Build Model

- Always start with the simplest model
 - Linear Regression, Logistic Regression, ...
- Easy to debug
- Run fast, iterate quickly
- Good baseline for future work
- Better understanding of the problem and data

In HW4, we provide a project-starter code that guide you to build a simple linear regression model.