CSCI-SHU 360 Machine Learning: Final Project Proposal

Team members:

- *[Xingji Jax Li (xl2860)]*

Short description of the project: *[What is the topic? What is the main goal of the project?]*

Basically, the project is an ongoing Kaggle competition (link), which is about sales prediction for the thousands of product families sold at Favorita stores located in Ecuador.
I believe that it can be a perfect fit for the course project since we have covered several possible methods to tackle the issue and the provided data is at a suitable level that would not be too obvious to modeling but also too big or complex for my laptop to process effectively. And I would like to set the goal for this project as reaching a good prediction accuracy, ideally getting a high position in the competition leaderboard.

Data set: *[What kind of data do you need? How do you plan to get it? Do you already have some sources?]*
The data is provided in the competition link, where there are 7 csv files available and a few additional information that could be useful. Here are the basic description for each file and its entities:
Train.csv:

- main file used for training, in this case I would divided it into local train, validation and test dataset
- Id: the index for sell history in the form data-store-family-sales-onpreomotion
- Date: date range in 2013-01-01 to 2017-08-15
- Store_nbr: index for stores
- Family: categories for the products, 33 kinds in total
- Sales: total sales for a product family at a particular store at a given date
- Onpromotion: gives the total number of items in a product family that were being promoted at a store at a given date

Test.csv:

- Same entities as train.csv except the sales (to predict)
- Date: date range 2017-08-16 to 2017-08-31, 15 days after the train.csv data

Stores.csv:

- Addition information for stores, entities are mostly categorical
- Store_nbr: store index
- City: city of the store place
- State: state of the store place
- Type: store type
- Cluster: a grouping of similar stores

Oil.csv:

- Daily oil price. Includes values during both the train and test data timeframes. (Ecuador is an oil-dependent country and its economic health is highly vulnerable to shocks in oil prices.)

Holidays_event.csv:

- Holidays and Events, with metadata
- Additional Note: Pay special attention to the **transferred** column. A holiday that is transferred officially falls on that calendar day, but was moved to another date by the government. A transferred day is more like a normal day than a holiday. To find the day that it was actually celebrated, look for the corresponding row where type is Transfer. For example, the holiday Independencia de Guayaquil was transferred from 2012-10-09 to 2012-10-12, which means it was celebrated on 2012-10-12. Days that are type Bridge are extra days that are added to a holiday (e.g., to extend the break across a long weekend). These are frequently made up by the type Work Day which is a day not normally scheduled for work (e.g., Saturday) that is meant to pay back the Bridge.
- Additional holidays are days added to a regular calendar holiday, for example, as typically happens around Christmas (making Christmas Eve a holiday).

Transactions.csv:

- For each date-store, there is a feature named transactions, not sure what it means.

Sample_submission.csv:

- A example for the format of submission

Additional information:

- Wages in the public sector are paid every two weeks on the 15 th and on the last day of the month. Supermarket sales could be affected by this.
- A magnitude 7.8 earthquake struck Ecuador on April 16, 2016. People rallied in relief efforts donating water and other first need products which greatly affected supermarket sales for several weeks after the earthquake.

Project plan: *[What will be the main steps to develop your project? What kind of techniques do you plan to use? This does not need to be your final plan, but the earlier you have a plan, the earlier we can provide feedback and support.]*
Basic steps and timeline

Stage	Task	Task Breakdown	Ideal due	Result & Note
Proposal	Proposal version 1		11.04.2022	This file
Exploratory Data Analysis (EDA)	Possible sales segregation by store	By city	11.14.2022	
		By state		
		By type		
		By group		
	Sales correlation with holliday			
	Sales correlation with oil price data			
Feature Choosing and Engineering	Choosing and normalization based on EDA		11.15.2022	
Modeling	Regression	Linear	11.20.2022	
		Poly		
		With penalty(LASSO or Ridge)		
	Decision Tree		11.25.2022	
	Others learnt in the future			
	Model choosing and combining		11.30.2022	
Post-Model engineering	Use sub models			
	Total lift?			
Finalized Model			12.8.2022	
Submission & Check Performance			12.10.2022	
Presenting			12.15.2022	