Springboard—Data Science Career Tack
Grocery Store Sales Forecasting
Capstone Project 1 Proposal
By Henry Ruan
February 2020

My first capstone project will be based on the Kaggle competition project: “Corporacion Favorita Grocery Sales Forecasting.” Grocery store sales forecasting is a critical factor for store managers to growth their sales revenue, maintain a healthy level of inventory and improve their operation profit.

In this project, I’ll consider the application of several data science tools and machine learning algorithms and Python packages, such as linear regression and **ARIMA** to build models to predict the unit sales for thousands of items sold at different stores of this company.

I am planning on considering various business activities and operational aspects that might have an impact on predicting the store sales, for example store location, available inventory, pricing, operation hours, holiday/events and promotion on specific type of products/items, etc.

I am planning on studying approaches to address this project by considering  the following three layers:

- First level is to predict the sales on typical products/items at one grocery store/location
- Second level is to predict the sales on typical products/items at multiple grocery stores/locations
- Third level is to predict the sales on typical products/items at multiple grocery stores/locations on different periods (daily, weekly, monthly, quarterly)

The overall approach taken to solve the problem is to first build familiarity with the data and business domain through descriptive statistics and data visualization, and then use this understanding to develop a set of hypotheses to explain the increase in sold items experienced by the stores.

Some hypothesis to consider are:

Hypothesis 1: Type of products sold are dependent on territory/holiday/events/seasonal patterns

Hypothesis 2: The number of units to be sold are highly dependent on the pricing, availability.

Hypothesis 3: Certain grocery stores reload the inventory more frequently.

In practice, to get the predictive models implemented, there will be a potential improvement on the overall management of groceries and on their activities among the stakeholders: customers, nearby community, logistics/supply chain, sales/marketing, management, operation & financing, competitors.

The data sets are available at https://www.kaggle.com/c/favorita-grocery-sales-forecasting/data. These data sets seem to be already mostly clean and saved as CSV files.

The training dataset includes dates, store and item information, whether that item was being promoted, as well as the unit sales. Additional files include supplementary information that may be useful in building the models:

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

My project plan and schedule in alignment with curriculum and requirements, as suggested by my mentor, and with deadlines estimated by me, is as follows:

(1) Complete and submit capstone project proposal, per unit 4.3

(2) Complete and submit the Data Wrangling notebook+report after completing all mini-projects in unit 5 of the curriculum

(3) Complete and submit the Data Storytelling notebook+report, per unit 7

(4) Complete and submit the application of Inferential Statistics notebook+report to my CP dataset after finishing all mini-projects in unit 8 of the curriculum

(5) Complete and submit the Milestone Report for my CP
See unit 8.5 of the curriculum.

(6) Build a baseline model notebook for your CP using "pmdarima”

(6.1) To assess the performance of my model/s, do the following:

(a) Split the dataset into training and test set, determining an appropriate split. For time series forecasting, it is customary for the training set to be in the chronological past, and for the test set in the chronological future of the training set. (b) Build time series model/s using the training set.

(c) Compute RMSE, MAPE (*) and R^2 for the training set.
This might be implemented in different ways depending on the package you are using (see "Resources" below).

(d) Compute RMSE, MAPE and R^2 for the test set.

(e) Determine if there is overfitting, by comparing (c) and (d).

(f) Show the following graphs:

(f.1) Predicted vs Actual scatterplot.

(f.2) Predicted vs Residuals scatterplot.

(f.3) Histogram of Residuals, identifying positive and negative residual worst-cases for 90% and 95% of the histogram density.

(6.2) Present an analysis of the performance of the model/s in light of (6.1).

(7) Discuss with my mentor applicable extensions to (6).

Options include (see "Resources" below):

- Prophet.
- Deriving features to apply a regression model, and ...
- H2O.

(8) Implement the extensions agreed upon in (7).

(9) Prepare a first draft of CP Final Report, per the Capstone Project  Guidelines (CPG) document, and submit it to my mentor for review **Due on 5/28/2020**

(10) Use feedback from my mentor after (9) to generate the CP Final Report, Presentation Slide Deck, and cleaned-up code, and submit them as deliverables for your CP.

=================================================================
(*) Resources provided by my mentor:
=================================================================
(0) This reference is well-known and respected and covers the theory of ARIMA-type of models: https://otexts.com/fpp2/

Although the examples are in R, the explanations are language-agnostic.
Most of the approaches presented here require manual intervention, which is why the approaches below would be appropriate when automating the forecasting with many time series.

(1) pmdraima (formerly Pyramid) is a Python implementation of auto-arima (maybe discussed in fpp above?). Some associated resources are:
https://github.com/tgsmith61591/pmdarima
https://www.alkaline-ml.com/pmdarima/

Even more resources can be found by searching pyramid python time series

(2) Prophet was developed by Facebook:
https://facebook.github.io/prophet/
https://towardsdatascience.com/a-quick-start-of-time-series-forecasting-with-a-practical-example-using-fb-prophet-31c4447a2274

More resources ... search for prophet time series

(3) More recent approaches incorporate the concept of Auto-ML. An example is H2O ...
https://www.h2o.ai/blog/time-money-automate-time-series-forecasts-driverless-ai/
More resources ... search for h2o time series

(4) Time series forecasting as a supervised machine learning problem
https://machinelearningmastery.com/time-series-forecasting-supervised-learning/
https://machinelearningmastery.com/convert-time-series-supervised-learning-problem-python/

(5) Backtesting
(5.0) General:
https://medium.com/cindicator/backtesting-time-series-models-weekend-of-a-data-scientist-92079cc2c540
(5.1) pmdraima:
https://alkaline-ml.com/pmdarima/auto_examples/model_selection/example_cross_validation.html#sphx-glr-auto-examples-model-selection-example-cross-validation-py

(5.2) prophet:
https://blog.exploratory.io/a-gentle-introduction-to-backtesting-for-evaluating-the-prophet-forecasting-models-66c132adc37c