Capstone Proposal meshari aldossari

Machine Learning Engineer Nanodegree December 9th, 2021

Domain Background:

Sales forecasting is one of the first known uses in machine learning and it helped in machine learning acceptance in the market as its importance is clear from a business standpoint being able to predict the sales for the future would help especially in the logistic related problems being able to predict sales should have a noticeable help in them as logistic can be prepared ahead of time and be ready to use.

not only is Sales forecasting important in the logistic department, but also in the finance department since it can help in knowing your approximate income from these sales which would lead to better financial related decisions.

Many companies reward employees for performance against forecast (or more precisely, targets which are derived from forecasts), using an asymmetric system where exceeding the forecast results in positive rewards whereas falling short results in a mixture of punishments and withholding of rewards [1].

problem statement:

the goal is to use time-series forecasting to forecast store sales on data from Corporación Favorita, a large Ecuadorian-based grocery retailer.

Current subjective forecasting methods for retail have little data to back them up and are unlikely to be automated. The problem becomes even more complex as retailers add new locations with unique needs, new products, ever-transitioning seasonal tastes, and unpredictable product marketing.

datasets and inputs:

the dataset consists of six files:

- train.csv: The training data, containing time series of features store_nbr, family, and onpromotion as well as the target sales.
- test.csv: The test data, having the same features as the training data.
- sample_submission.csv: A sample submission file in the correct format.
- stores.csv: Store metadata, including city, state, type, and cluster.
- 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_events.csv: Holidays and Events, with metadata

The data is available as a Kaggle competition you can access the data from the link below to get a more detailed view of data specifications:

https://www.kaggle.com/c/store-sales-time-series-forecasting/data

Solution Statement:

As the problem is a regression problem, so I would use supervised regression algorithms, such as Linear Regression, Support Vector Regressor, Random Forest Regressor, Lasso Regression, Decision Tree, and Gradient Boosting.

I would also test out a few models designed for time series forecasting and compare it's results with the known supervised regression algorithms mentioned above.

After testing a few of those models, I would choose the one with highest score

benchmark model:

I would use the simple model Linear Regression as a benchmark model.

evaluation metrics:

The evaluation metric for this competition is Root Mean Squared Logarithmic Error.

The Squared Logarithmic Error is calculated as:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(y_i+1)-\log(\hat{y}_i+1))^2}$$

where:

- n is the total number of instances.
- $\hat{y}_i$ is the predicted value of the target for instance (i).
- y_i is the actual value of the target for instance (i).
- log is the natural logarithm.

project design:

- step 1: explore data.
- step 2: preprocess data.
- step 3: choose a model.
- step 4: test the model result.
- step 5: select the model with highest score.

this a simple design in which many details wasn't discussed, but this is the general flow for this project, in which step 3, and step 4 would be repeated to test out a few different models.

In the preprocessing step I would try to come to a way to represent the time steps using the date provided, also to merge the different files data and make use of them in making the correct prediction e.g., Adding oil price for each entry based on its sales date.

References:

1. Fildes, Robert, et al. "Researching sales forecasting practice: Commentaries and authors' response on "Conducting a Sales Forecasting Audit" by MA Moon, JT Mentzer & CD Smith." International Journal of Forecasting 19.1 (2003): 27-42.