Case Study : M5 Forecasting - Accuracy

Overview of Case Study

Introduction

This competition is organized by Makridakis Open Forecasting Center (MOFC) at the University of Nicosia which aims at advancing the theory and practice of forecasting.In this competition we are provided with the sales data on 3049 products of Walmart for over a period of 5 years (from 29/01/2011 to 23/05/2016).Here we are needed to forecast sales of these products for the upcoming 28 days using the given data and machine learning.

Business Problem

The problem given here is to forecast the sale of the product in the future days.This problem tries to find the factors that ensures smooth functioning of the business by creating customer centric experience by making available more in demand products readily and always ,which in turn will boost revenue of the organization.Also the fulfillment lag between the demand and supply will be identified by knowing beforehand that what quantity of product will be sold in future which will ensure customer satisfaction .

Also staff efficiency will also be improved as they will be assisted with this forecasting system for better allocation of workforce and goods to various store locations in different states, as equal distribution of products is not very optimized but if done according to demands of that area will save lot of logistics cost and also reduce wastage of product stagnating at places where not required in proportion to the demand.

So in conclusion the stakeholders here in this problem are store managers,logistics department,customer and store owner and if only the demand is known then only a efficient store productivity for managers,reduce logistics sfaction and increase revenue for owner by reducing waste and optimizing business processes and introduce predictability in revenue growth and sales.

ML problem formulation

Time-series forecasting and Regression To find number of each products that will be sold each day in each store from the date of 24/05/2016 to 19/06/2016 , given the date ,day information like weekday,event happening around the area , special discount event like SNAP ,product price and event type. To solve the above we would be using data collected in Jan 2011 to May 2016 for predicting the sales of all the products in coming 28 days from the date of 24/05/2016.

Business constraints

No strict latency concerns Impact of inaccuracy is low on business persay , but high accuracy is needed as per competition's demand

Dataset analysis

The dataset involves the unit sales of 3,049 products, classified in 3 product categories (Hobbies, Foods, and Household) and 7 product departments, in which the above-mentioned categories are disaggregated. The products are sold across ten stores, located in three States (CA, TX, and WI). In this respect, the bottom-level of the hierarchy, i.e., product-store unit sales can be mapped across either product categories or geographical regions.Below diagram gives the aggregation level of products based on above categories.

Performance Metrics

In this problem I would like to use RMSE and MAE as these two metrics commonly used for time series forecasting and try a custom metric for this problem UMBRAE from this [link]

Solutions/Research Papers/Kernels

Solutions

First Solution: Lokesh Gupta

Link : https://www.kaggle.com/anshuls235/time-series-forecasting-eda-fe-modelling

In this kernel of kaggle the author followed follwing methods

Saved successfully!

% to avoid ram crashing.

ong form and merging all three csv files into one to form trainable dataset

3. Performed basic time series feature engineering to come up with feature like : Label Encoding , Lag , Mean encoding , Rolling window stat , Expanding window stats
4. Data modelling is done using LighGBMRegressor

Second Solution: Shahrukh Sharif

Link : https://github.com/ShahrukhSharif/m5_demand_Forecasting

In this github repo the author followed following methods

1. All the techniques in above techniques are follweda as in first solution.
2. Few new features are introduced in addition like : date related features(week,month,quarter,day of year).
3. Features related to special events were also included in the feature engineering.
4. Data is trimmed down and used from d-1050 for reducing computational requirement.
5. Trained Simple Moving Averages, ExtraTressRegression, RandomForestRegression & LightGBM and performed hyperparameter tuning.
6. The best model found was LightGBM.

Third Solution: Belinda Trotta

Link : https://www.kaggle.com/c/m5-forecasting-accuracy/discussion/163154

Metric : RMSE . WRMSSE evaluation metric is noisy, especially for features with short history, because random fluctuations in the day-to-day sales history can cause products to be weighted very differently even if they have similar long-term average. The approach used here is to train separate model to predict each day of forecasting horizon and creating features according to the same.The author went for this approach as according to a discussion in competition forum the recursive approach was coincidentally performing well on the training period

Features are made using the data of 3 years but training is done using only one year with the exclusion of the month december.These include lagged sales at various levels of aggregation with top down approach to keep the data more generalist with a pattern involved being less noisy.

This approach helps to focus on the effect of organizer provided custom metric WRMSSE which is pointed out to be noisy so I will not use the given metrics in my solution also lagged feature at multiple level of aggregation was very effective in order to get better prediction, so in feature engineering i will be going to use the different aggregate time series and try multiple combination to get best result.

Saved successfully!

orecasting-accuracy/discussion/163216

In this solution the idea is to train separate models for each 7 day of the forecast horizon for which the sales numbers of those weeks across the training set will be taken as the dataset for the respective models. Features used were very general i.e. time series features , price features,calendar feature,and no use of recursive feature.

Author also gives up on the suggested performance metrics WRMSSE and moves forward with rigorous cross validation by choosing 5 set of forecast horizon that is (d1578-d1605, d1830-d1857, d1858-d1885, d1886-d1913, d1914-d1941) with regular error metrics RMSE.

The model here used was LGBM with objective function being tweedie The general idea in the whole approach of the author is to use general ,easily reproducible features which are more fitting to practical grounds.And not a model which is only specific to the problem in hand.

In this solution the emphasis is on building separate models for subsets of the forecasting horizon and performing cross validation with different multiple test forecasting horizons. Hence the use of multiple cross validation with different sets of values will help in getting solid hyperparameter tuning

Fifth Solution : YeonJun In

Link: https://www.kaggle.com/c/m5-forecasting-accuracy/discussion/163684

This approach is trying to use two methods of modelling ,that are recursive and non-recursive individually but after training and looking at his metrics the author concluded that he would not use a single method as both methods tend to introduce uncertainty on cross validation of forecast.

Therefore the author chose to use both methods in his solution by ensembling them.Hence to further improve my solution i will explore the best of both the recursive and the non recursive methods and come up with the best forecasting method.

My first Cut approach

- First we need to downcast the data in to lower byte format to avoid ram crashing
- Second we preprocess the data into multiple features like like date feature,event feature,moving average,expanding average,and price features.
- Third we will choose train data and also multiple cross validation forecasting horizons.
- Fourth we will will train our model on two method first with recursive feature and second with non recursive feature and build a baseline model In this section I would like to implement a research paper proposing a custom model maked as rectify from this paper

Saved successfully!

ent are
ost,Facebook Prophet

- Fifth I would like to keep my training data just enough to get an acceptable accuracy and for which i will only consider data after 2 years as their wont be much impact of data of 3 years back on the forecating horizon
- Sixth i would like to training model on three ways

1. Separate model for each day of forecasting horizon
2. Separate model for each week of forecating horizon
3. Single model for the entire forecating horizon.

Reference

- https://stats.stackexchange.com/questions/389291/strategies-for-time-series-forecasting-for-2000-different-products
- https://machinelearningmastery.com/time-series-forecasting-performance-measures-with-python/
- https://www.linkedin.com/pulse/how-choose-quality-metric-forecasting-time-series-mape-zavalich/
- https://robjhyndman.com/papers/rectify.pdf :on custom modelling method
- https://dzone.com/articles/lessons-learnt-while-solving-time-series-forecasti-1
- https://arxiv.org/pdf/1811.10192.pdf : on tweedie gradient boosting for zero inflated data
- https://www.kaggle.com/c/m5-forecasting-accuracy/discussion/143070 : Making custom metric
- https://arxiv.org/pdf/1912.00370.pdf : Hierarchical time series forecasting
- https://www.artefact.com/news/sales-forecasting-in-retail-what-we-learned-from-the-m5-competition-published-in-medium-tech-blog/
- https://machinelearningmastery.com/multi-step-time-series-forecasting/
- https://machinelearningmastery.com/* * multi-step-time-series-forecasting-with-machine-learning-models-for-household-electricity-consumption/
- https://otexts.com/fpp2/accuracy.html

Saved successfully!