

M5 Forecasting - Accuracy: Estimate the unit sales of Walmart retail goods

Flavio Amurrio-Moya
George Mason University
4400 University Dr, Fairfax, VA 22030
famurrio@gmu.edu

Pyoungh Kang Kim
George Mason University
4400 University Dr, Fairfax, VA 22030
pkim23@gmu.edu

Abstract

Methods based on decision trees dominate the field of tabular regression problems and have shown to outperform many other methods such as weighted averaging and other statistical methods and even outperforming methods that are natively time series based on using networks such as LSTM/GRU. In this report, we plan to test out a pure Deep Learning approach which leverages key foundation concepts in both supervised and unsupervised learning to attempt to reach an acceptable score with minimum data manipulation, and virtually no data gathering/wrangling.

The “M5 Forecasting - Accuracy: Estimate the unit sales of Walmart retail goods” competition (will be referred to as the “M5 Forecasting Competition” and “The Competition”) challenged Kaggle users to design a learning model that would be able to predict how many of a certain item would be sold on a given day. In this paper, we embark in the challenge to create such a model and improve on current strategies.

1. Introduction

The M5 Forecasting competition was created on Kaggle by “The Makridakis Open Forecasting Center (MOFC) at the University of Nicosia”. This was done with the intent of being able to forecast the sales of items in order to minimize opportunity loss (such as not having enough of an item in stock) as well as to avoid stocking too much of an unpopular item. This competition aims to achieve more accurate and better-calibrated forecasts, reduce waste and be able to appreciate uncertainty and its risk implications.

In this paper, we discuss an approach to tabular regression learning that is an end to end Deep Learning solution. We try various modifications to the network and its loss function and discuss key points that will help achieve higher scores on the Competition.

We were provided with hierarchical sales data from Walmart to forecast daily sales for the next 28 days. The data covers stores in three US states (California, Texas, and Wis-

consin) and includes item level, department, product categories, and store details. It also contains explanatory variables such as price, promotions, day of the week, and special events.

1.1. Data

We were given a star schema dataset of csv files that contained 3 files to be used for training.

1.2. Closer Look At The Data

In Figure ?? os

In Figure ?? os

In Figure ?? os

In Figure ?? os

2. Approach

The approach corresponds with the files attached to the BlackBoard submission, which contains all the source code used to generate files, train, and run evaluation on.

2.1. Dataset

In order to train a Neural Network that can learn to correctly predict the number of sales per the hierarchical data shown above, i.e. item in a department, in a particular store, in a particular state, for a particular range of dates, we collapsed the star schema into a singular table containing all this particular information, with the dates and corresponding sales both becoming respondent columns. The training file ended up looking as such.

sales.csv - training file for the neural network, containing labels and its respective metadata. schema -

The column sales correspond to the number of products sold at that certain date per the hierarchical item. In order to develop such a training set, it is necessary to check for the correctness of the data. We made sure there are no duplicate columns/rows, and that all data was filled correctly for a correct merge operation between the respective fields across the tables. All the joins were done using outer joins, so

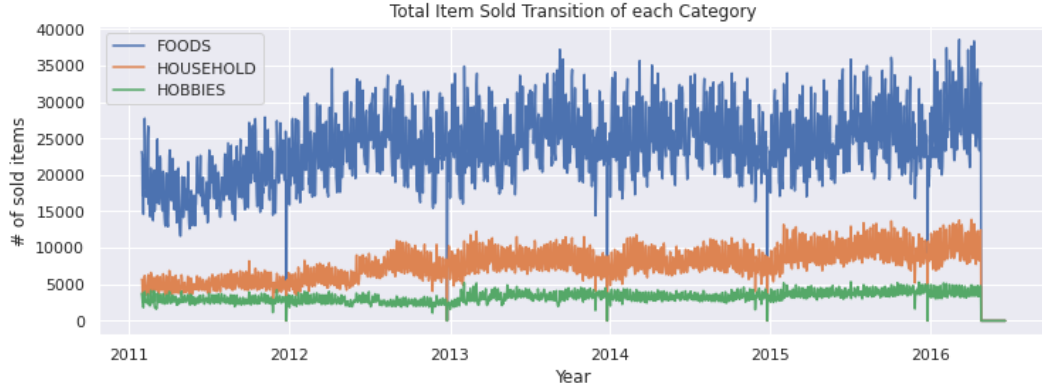


Figure 1. Total items sold of each category.

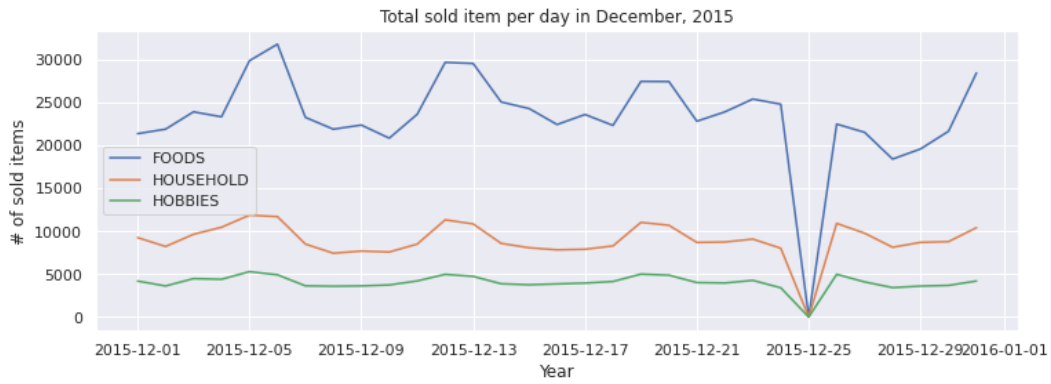


Figure 2. Total items sold per day in December 2015.

come resulting rows had missing fields such as a pricing per item at a particular date

sales.csv having rolled out the dates into columns of their own, and having merged with its surrounding tables, it became over 60 million rows, a job for a server computer with at least 50gb of ram. At the time, we didn't have access to a server computer, thus partitioned data and then finally concatenated them together. Once we had access to a server computer with 128gb of ram, we were able to load the dataset into memory. Loading the file into a pandas dataframe took over 4 minutes. However, after manually typing the pandas' columns to smaller data types, the load took somewhere around 3-4 minutes. The type was defined as below.

2.2. Model/Data

The model we used was a tabular model from a popular repository on GitHub, <https://github.com/fastai/fastai>. The underlying details worked as so. We defined the dataset to be fed to the model as a TabularDataBunch, which is a hierarchical abstraction to the DataLoader and Dataset classes

and manipulation as given in PyTorch for tabular datasets. It bunches the data up to be trained upon and customized upon, as well as preprocessing. Our sales.csv confirmed to their API and was loaded successfully. The DataBunch API allowed us to fill in the previously mentioned data such as pricing, by using the median of the values across the column. We also normalize the pricing column using Standardization. The prices ranged from .01 to 107.32, before Standardization.

$$Z = \frac{x - u}{\phi} \quad (1)$$

Z = standard score

x = observed value

u = mean of the sample

ϕ = standard deviation of the samples

The model uses embeddings under the hood in order to learn meaningful vectors for categorical variables during training, that can be referenced during evaluations. Then

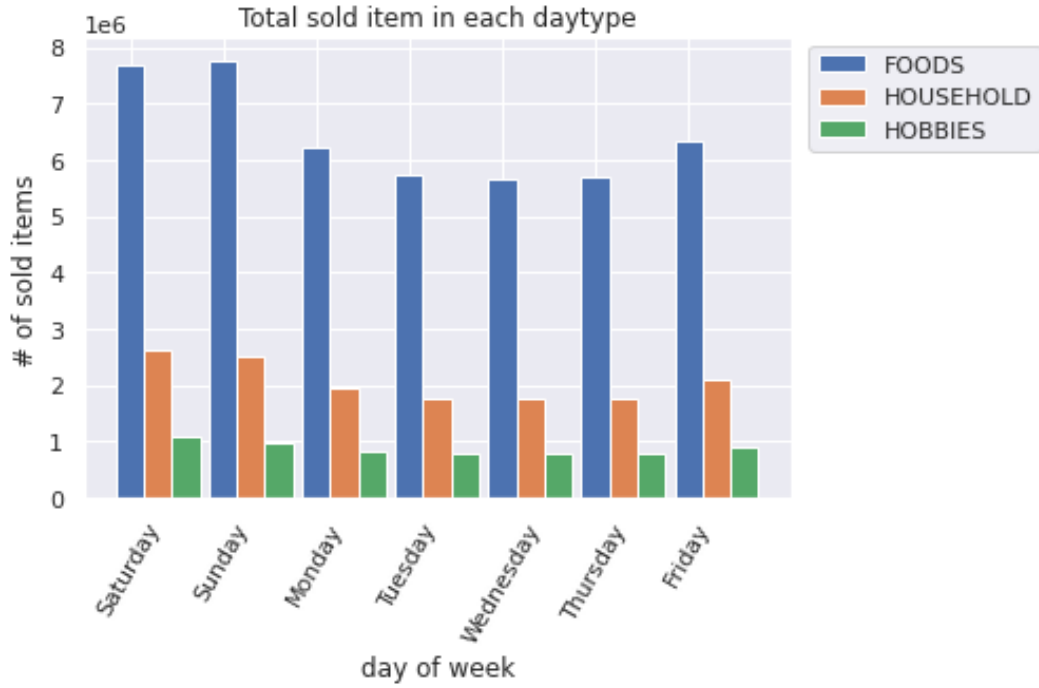


Figure 3. Total items sold each day of the week

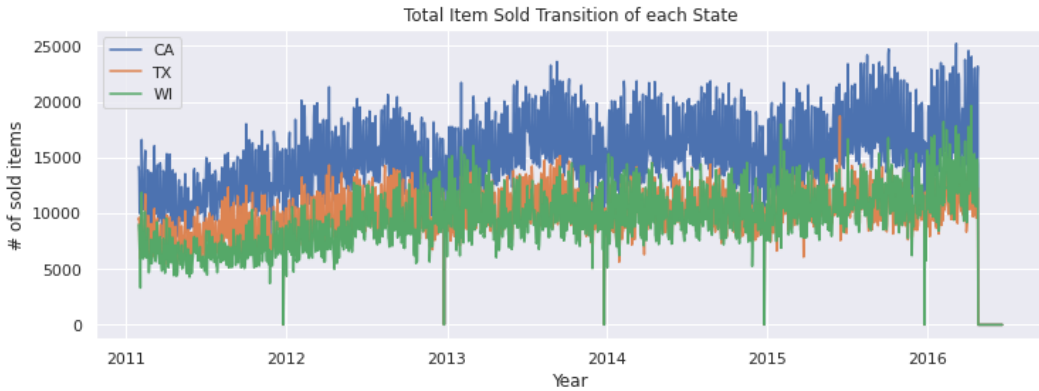


Figure 4. Total items sold in each state.

subsequent layers are defined to learn the interaction between these embeddings and the dependent variable, i.e. sales. It contains BatchNorm layers, to speed up training as well as adding an extra layer against overfitting, and a ReLU layer for nonlinearity. The dropout rate can be specified for both embeddings and the subsequent layers.

Embeddings can have a vector size associated with them, and this is defined from a heuristic that the authors seem to have come up with through empirical means. The embeddings also have an extra size to their vocab(number of categories), in order to have another vector for an unknown field that is encountered, i.e. na.

Below is a view of the model of the best results on our Kaggle submissions, as we have tried increasing the layers on the model to no avail, perhaps due to the limited training time we had. Again, the Embeddings are for our categorical variables + na, whereas our only continuous variable 'sell_{prices}' are fed in as the normalized value directly. The method to do it is:

2.3. Loss Functions/Forward Methods

There were two different options for the Forward functions. We tried both.

1. A simple output through the layers of the neural net-

work, with no modifications to the output variable.

2. A squashed and re-expanded output by means of using the sigmoid function with a specified dependent variable range.

A simple output through the layers of the neural network, with no modifications to the output variable.

$$(y_1 - y_0) * \text{sigmoid}(x) + y_0 \quad (2)$$

This method is particularly useful when you are more interested in ratios of the output being close to each other, especially when paired with a method of logging the dependent variable and inverting the logging for submission. However, the mean(1.0312) and scale(0-763) of the dependent was not appropriate for this forward pass. The intuition is that sigmoid tends to saturate towards 0 and 1, and having outliers that raises the scale makes it a harder task for the Neural Network. The results, as will be discussed in the Results/Extras section shows this.

The loss for both the forward passes is a mean squared error loss function.

$$MSE = \frac{1}{n} \sum n(y_i - \hat{y}_i)^2 (y_1 - y_0) * \text{sigmoid}(x) + y_0 \quad (3)$$

This is typical of regression tasks which tries to minimize the residual sum of squares.

The metric used in the paper is Weighted Root Mean Squared Scaled Error (RMSSE), however in terms of training, it's fine to use a MSE.

2.4. Training

Training the Neural Network can be a tedious task, with iterative and manual methods to try and converge the train/valid/test losses down to zero. However, we took a much more scientific approach as given from the API of FastAi. We used a learning rate finder algorithm, which employs a callback while increasing the learning rate, to see when the losses would diverge. Then we take the learning rate right before the diversion, and scale it down by a factor of 10-100 and choose that as the learning rate.

For the actual SGD process, we used a fit one-cycle method on top of the Adam Optimizer. The adam optimizer from PyTorch were given betas of (.95, .99) for the calculation of moving averages of the gradients, and the default epsilon value of 1e-08 to the denominator for numerical stability.

3. Results

The results get explained here.

4. Related Work

The relate work goes here

5. Summary/Discussion/Conclusion

The Summary Goes here

6. References

The References Goes here

7. \LaTeX Reference Section

`\cvprfinalcopy` Is for different font
(e.g. this line is 095.5)

An example of a bad paper:

An analysis of the frobnicatable foo filter.

In this paper we present a performance analysis of our previous paper [1], and show it to be inferior to all previously known methods. Why the previous paper was accepted without this analysis is beyond me.

[1] Removed for blind review

This is how you cite a reference submission [?] as additional material and

and this is a quote and a another special formatting `eccv06.pdf`.

require more special formatting

“Zero-g frobnication: How being the only people in the world with access to the Apollo lander source code makes us a wow at parties”, by Zeus *et al*.

FAQ: Are acknowledgements OK? No. Leave them for the final copy.

Compare the following:

`$conf_a$` *conf_a*

`conf_a` *conf_a*

See The \TeX book, p165.

The space after *e.g.*, meaning “for example”, should not be a sentence-ending space. So *e.g.* is correct, *e.g.* is not. The provided `\eg` macro takes care of this.

When citing a multi-author paper, you may save space by using “et alia”, shortened to “*et al.*” (not “*et. al.*” as “*et*” is a complete word.)

For this citation style, keep multiple citations in numerical (not chronological) order, so prefer [?, ?, ?] to [?, ?, ?].

This is fractions text area is $6\frac{7}{8}$ inches (17.5 cm) wide by $8\frac{7}{8}$ and more $\frac{5}{16}$ and other 8.5×11 -inch

Reference figures Figures ?? and ??. Short captions should be centred.

For no indent

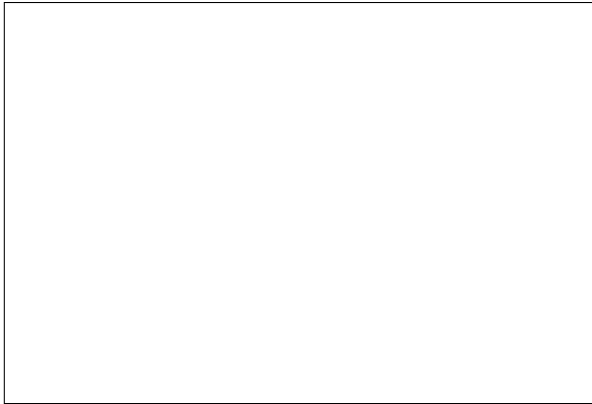


Figure 5. Example of caption. It is set in Roman so that mathematics (always set in Roman: $B \sin A = A \sin B$) may be included without an ugly clash.

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Table 1. Results. Ours is better.

FIRST-ORDER HEADINGS. (For example, **1. Introduction**)

SECOND-ORDER HEADINGS. (For example, **1.1. Database elements**)

Please use footnotes¹ sparingly.

When placing figures in \LaTeX , it's almost always best to use `\includegraphics`, and to specify the figure width as a multiple of the line width as in the example below

```
\usepackage[dvips]{graphicx} ...
\includegraphics[width=0.8\linewidth]
{myfile.eps}
```

¹This is what a footnote looks like. It often distracts the reader from the main flow of the argument.