**Capstone Project** **Document**

Prediction of future sales in Walmart

Dong Qin

# Project overview

Accurate **sales predictions** enable companies to make informed business decisions and **forecast** short-term and long-term performance. **Sales predictions** also gives insight into how a company should manage its workforce, cash flow, and resources.

In this project, we aim to predict the future sales in Walmart based on the purchase history. We first analyse and explore the basic information of the provided dataset, and then extracted useful features. Based on the extracted features, we build various regression models to predict future sales. In order to improve the regression performance, all these models are ensembled via stacking.

# Project goal

Overall goal of this project is to estimate, as precisely as possible, the point forecasts of the unit sales of various products sold in the USA by Walmart. Specifically, we use hierarchical and historic sales data from Walmart, the world’s largest company by revenue, to forecast daily sales for the next days.

# Project values

* to advance the theory and practice of forecasting.
* be applied in various business areas, such as setting up appropriate inventory or service levels.
* help distribute the tools and knowledge so others can achieve more accurate and better calibrated forecasts
* reduce waste and be able to appreciate uncertainty and its risk implications.

# Stakeholders

Business manager, technic manger, team member and salesmen.

# Data basics

The M5 dataset, generously made available by Walmart, involves the unit sales of various products sold in the USA, organized in the form of grouped time series. More specifically, 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, as follows:

**Table 1: Number of M5 series per aggregation level.**

|  |  |  |
| --- | --- | --- |
| **Level**  **id** | **Aggregation Level** | **Number of series** |
| **1** | Unit sales of all products, aggregated for all stores/states | 1 |
| **2** | Unit sales of all products, aggregated for each State | 3 |
| **3** | Unit sales of all products, aggregated for each store | 10 |
| **4** | Unit sales of all products, aggregated for each category | 3 |
| **5** | Unit sales of all products, aggregated for each department | 7 |
| **6** | Unit sales of all products, aggregated for each State and category | 9 |
| **7** | Unit sales of all products, aggregated for each State and department | 21 |
| **8** | Unit sales of all products, aggregated for each store and category | 30 |
| **9** | Unit sales of all products, aggregated for each store and department | 70 |
| **10** | Unit sales of product *x*, aggregated for all stores/states | 3,049 |
| **11** | Unit sales of product *x*, aggregated for each State | 9,147 |
| **12** | Unit sales of product *x*, aggregated for each store | 30,490 |
| **Total** | | **42,840** |

A close up of a logo

Description automatically generated

**Figure 1: An overview of how the M5 series are organized.**

The historical data range from **2011-01-29** to **2016-06-19**. Thus, the products have a (maximum) selling history of 1,941[[1]](#footnote-1) days / 5.4 years.

The data covers stores in three US States (California, Texas, and Wisconsin) and includes item level, department, product categories, and store details. In addition, it has explanatory variables such as price, promotions, day of the week, and special events. Together, this robust dataset can be used to improve forecasting accuracy.

The M5 dataset, generously made available by Walmart, involves the unit sales of various products sold in the USA, organized in the form of grouped time series. More specifically, the dataset involves the unit sales of \*\*3,049 products\*\*, classified in \*\*3 product categories\*\* (Household, Foods, and Hobbies) and \*\*7 product departments\*\* (

HOUSEHOLD\_1, HOUSEHOLD\_2, FOODS\_1, FOODS\_2, FOODS\_3, HOBBIES\_1, HOBBIES\_2), in which the above-mentioned categories are disaggregated.

The products are sold across ten stores, located in \*\*3 States\*\* (CA, TX, and WI) and \*\*10 shops/stores\*\* (CA\_1, CA\_2, CA\_3, CA\_4; TX\_1, TX\_2, TX\_3; WI\_1, WI\_2, WI\_3).

The historical data range from 2011-01-29 to 2016-06-19. Thus, the products have a (maximum) selling history of 1,941 days / 5.4 years.

Next, we introduce two files we used in this project:

**File 1: “calendar.csv”**

Contains information about the dates the products are sold.

date: The date in a “y-m-d” format.

wm\_yr\_wk: The id of the week the date belongs to.

weekday: The type of the day (Saturday, Sunday, …, Friday).

wday: The id of the weekday, starting from Saturday.

month: The month of the date.

year: The year of the date.

event\_name\_1: If the date includes an event, the name of this event.

event\_type\_1: If the date includes an event, the type of this event.

event\_name\_2: If the date includes a second event, the name of this event.

event\_type\_2: If the date includes a second event, the type of this event.

snap\_CA, snap\_TX, and snap\_WI: A binary variable (0 or 1) indicating whether the stores of CA, TX or WI allow SNAP purchases on the examined date. 1 indicates that SNAP purchases are allowed.

**File 2: “sales\_train.csv”**

Contains the historical daily unit sales data per product and store.

item\_id: The id of the product.

dept\_id: The id of the department the product belongs to.

cat\_id: The id of the category the product belongs to.

store\_id: The id of the store where the product is sold.

state\_id: The State where the store is located.

d\_1, d\_2, …, d\_i, … d\_1941: The number of units sold at day i, starting from 2011-01-29.

# Exploratory Data Analysis

Shape (30490, 1947)

No null values

Sales distributions on Weekdays (left) and Weekends (right)

A screenshot of a cell phone

Description automatically generated A screenshot of a cell phone

Description automatically generated

Sales on each month

A picture containing drawing, fence

Description automatically generated

Sales on each festival

A picture containing drawing

Description automatically generated

Sales on different snap

A screenshot of a cell phone

Description automatically generated

# Modelling

Based on the visualisation in the previous section, we can find month feature shows no obvious difference on sales. That is why we select the other features, i.e., weekdays/weekends, festivals, Snap and State/Category as our predictors.

For each feature, we tried different regressor models and select the best one which achieves the highest R-squared value. In order to improve the regression performance, all these models are ensembled via stacking. In stacking, the first level is the models we obtained based on features, and the second level is a linear regression model as shown in below:

A close up of a piece of paper

Description automatically generated

# Outcomes and conclusion

|  |  |  |
| --- | --- | --- |
| **Prediction Model** | **R-Squared Train** | **R-squared Test** |
| Weekdays (LinearRegression) | 0.831 | 0.656 |
| Weekends (LinearRegression) | 0.805 | 0.715 |
| Sporting (LinearRegression) | 0.542 | 0.475 |
| Cultural (LinearRegression) | 0.714 | 0.675 |
| National (LinearRegression) | 0.617 | 0.600 |
| Religious (LinearRegression) | 0.697 | 0.687 |
| No-festival (LinearRegression) | 0.864 | 0.677 |
| Snap\_CA (LinearRegression) | 0.764 | 0.637 |
| Snap\_TX (LinearRegression) | 0.806 | 0.695 |
| Snap\_WI (LinearRegression) | 0.813 | 0.686 |
| No\_snap (LinearRegression) | 0.827 | 0.702 |
| Cat/State (**DecisionTreeRegression**) | 0.956 | 0.760 |
| **Stacking (LinearRegression)** | 0.858 | **0.845** |

* Linear Regression works well on numerical features
* Decision Tree Regression works well on categorical features
* Stacking combines weak predictors to achieve a high regression performance.

1. [↑](#footnote-ref-1)