Trabajo Final de Master

Predicción de Ventas de Comestibles Corporación Favorita

Nombre: Gabriel Kreplak
Plan: Master en Inteligencia de Negocio y Big Data
Área: Análisis de Datos

Nombre Consultora: Dra. Laia Subirats Maté
Nombre Profesora responsable de la asignatura: Dra. Teresa Sancho Vinuesa y Dra. María Pujol Jover

Fecha Entrega: 23 de enero de 2018

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Predicción de Ventas de Comestibles Corporación Favorita.*
Nombre del autor:	*Gabriel Kreplak*
Nombre del consultor/a:	*Dra. Laia Subirats Maté*
Nombre del PRA:	*Dra. Teresa Sancho Vinuesa y Dra. María Pujol Jover*
Fecha de entrega (mm/aaaa):	Enero 2018
Titulación::	*Máster Inteligencia de Negocio y Big Data*
Área del Trabajo Final:	*Análisis de Datos*
Idioma del trabajo:	*Español – Inglés*
Palabras clave	*Análisis de Datos, Ventas*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

El trabajo escogido tiene por finalidad conseguir una puntuación relevante en la competición de Kaggle: **Corporación Favorita Grocery Sales Forecasting**. El objetivo de la competición es construir un modelo predictivo para pronosticar ventas futuras.

Los métodos de planificación de ventas de esta compañía actualmente están poco sustentados en datos y como consecuencia no están automatizados. Por este motivo, Corporación Favorita propone a la comunidad Kaggle el desarrollo te técnicas de Machine Learning para conseguir adecuar en lo posible la logística y oferta de productos a la demanda de éstos donde son requeridos.

Como se detallará en el apartado de descripción de los ficheros de datos que se ofrecen para la predicción, éstos incluyen información para entrenamiento y test de los algoritmos supervisados y otros ficheros de datos para poder poner en contexto los datos mencionados. Los ficheros aportados son: train.csv, test.csv, simple_submission, stores.csv, ítems.csv, transactions.csv, oil.csv y holidays_events.csv .

El trabajo lo estructuraré en 2 partes principales:

a) Análisis Exploratorio de Datos, que incluirá que estudiará detalladamente la información de ventas: en general, por tipo de tienda, geolocalización , ventas por estado, por ciudad, ventas en el tiempo, correlación ventas – precio del petróleo, análisis de los productos y sus familias, transacciones, etc.

b) Estudio Predictivo para lo que utilizaré los algoritmos con los que obtenga un resultado que me permitan obtener una buena clasificación

en la tabla y recibir también comentarios positivos.

Abstract (in English, 250 words or less):

The chosen final grade project theme is to earn a relevant score in the Kaggle's competition: Corporación Favorita Grocery Sales Forecasting by building a state of the art predictive model aimed to forecast future sales.

Corporación Favorita has challenged the Kaggle community to build a model that more accurately forecasts product sales. They currently rely on subjective forecasting methods with very little data to back them up and very little automation to execute plans. They're excited to see how machine learning could better ensure they please customers by having just enough of the right products at the right time.

As it will be detailed below, in section "Description of the Input Data", training and testing data files are supplied in order to develop predictive supervised algorithms as well as other data file to be able to put the aforementioned data in context. The files provided are: train.csv, test.csv, simple_submission, stores.csv, items.csv, transactions.csv, oil.csv and holidays_events.csv.

The work will be structured in 2 main parts:

a) Exploratory Data Analysis, which will include detailed study of sales information: in general, by type of store, geolocation, sales by state, by city, sales in time, sales correlation - price of oil, analysis of products and their families, transactions, etc.

b) Predictive study for which I will use the state of the art algorithms in order to obtain a relevant classification in the score table and to receive positive comments as well.

Índice

Lista de figuras

No se encuentran elementos de tabla de ilustraciones.

1 Overview

Brick-and-mortar grocery stores are always in a delicate dance with purchasing and sales forecasting. Predict a little over, and grocers are stuck with overstocked, perishable goods. Guess a little under, and popular items quickly sell out, leaving money on the table and customers fuming.

The problem becomes more complex as retailers add new locations with unique needs, new products, ever transitioning seasonal tastes, and unpredictable product marketing. Corporación Favorita, a large Ecuadorian-based grocery retailer, knows this all too well. They operate hundreds of supermarkets, with over 200,000 different products on their shelves.

Corporación Favorita has challenged the Kaggle community to build a model that more accurately forecasts product sales. They currently rely on subjective forecasting methods with very little data to back them up and very little automation to execute plans. They're excited to see how machine learning could better ensure they please customers by having just enough of the right products at the right time.

2 Evaluation

Submissions are evaluated on the Normalized Weighted Root Mean Squared Logarithmic Error (NWRMSLE), calculated as follows:

$$NWRMSLE = \sqrt{\frac{\sum_{i=1}^{n} w_i \left(\ln\left(\hat{y}_i + 1 \right) - \ln\left(y_i + 1 \right) \right)^2}{\sum_{i=1}^{n} w_i}}$$

where for row i, $\hat{y}$ is the predicted unit_sales of an item and y_i is the actual unit_sales; n is the total number of rows in the test set.

The weights, w_i, can be found in the items.csv file (see the Data page). Perishable items are given a weight of 1.25 where all other items are given a weight of 1.00.

This metric is suitable when predicting values across a large range of orders of magnitudes. It avoids penalizing large differences in prediction when both the predicted and the true number are large: predicting 5 when the true value is 50 is penalized more than predicting 500 when the true value is 545.

3 Submission File

For each *id* in the test set, you must predict the *unit_sales*. Because the metric uses ln(y+1), submissions are validated to ensure there are no negative predictions.

The file should contain a header and have the following format:

```
id,unit_sales
125497040,2.5
125497041,0.0
125497042,27.9
etc.
```

4 Data Description

In this competition, you will be predicting the unit sales for thousands of items sold at different Favorita stores located in Ecuador. The training data 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 your models.

File Descriptions and Data Field Information:

4.1 train.csv

- Training data, which includes the target *unit_sales by date, store_nbr*, and *item_nbr* and a unique *id* to label rows.
- The target *unit_sales* can be integer (e.g., a bag of chips) or float (e.g., 1.5 kg of cheese).
- Negative values of *unit_sales* represent returns of that particular item.
- The *onpromotion* column tells whether that *item_nbr* was on promotion for a specified *date* and *store_nbr*.
- Approximately 16% of the *onpromotion* values in this file are *NaN*.
- NOTE: The training data does not include rows for items that had zero *unit_sales* for a store/date combination. There is no information as to whether or not the item was in stock for the store on the date, and teams will need to decide the best way to handle that situation. Also, there are a small number of items seen in the training data that aren't seen in the test data.

4.2 test.csv

- Test data, with the *date, store_nbr, item_nbr* combinations that are to be predicted, along with the *onpromotion* information.
- NOTE: The test data has a small number of items that are not contained in the training data. Part of the exercise will be to predict a new item sales based on similar products..
- The public / private leaderboard split is based on time. All items in the public split are also included in the private split.

4.3 sample_submission.csv

- A sample submission file in the correct format.
- It is highly recommend you zip your submission file before uploading!

4.4 stores.csv

- Store metadata, including *city, state, type*, and *cluster.*

- cluster is a grouping of similar stores.

4.5 items.csv

- Item metadata, including family, class, and perishable.
- NOTE: Items marked as *perishable* have a score weight of 1.25; otherwise, the weight is 1.0.

4.6 transactions.csv

- The count of sales transactions for each *date, store_nbr* combination. Only included for the training data timeframe.

4.7 oil.csv

- Daily oil price. Includes values during both the train and test data timeframe. (Ecuador is an oil-dependent country and it's economical health is highly vulnerable to shocks in oil prices.)

4.8 holidays_events.csv

- Holidays and Events, with metadata
- NOTE: Pay special attention to the *transferred* column. A holiday that is *transferred* officially falls on that calendar day, but was moved to another date by the government. A *transferred* day is more like a normal day than a holiday. To find the day that it was actually celebrated, look for the corresponding row where type is *Transfer*. For example, the holiday *Independencia de Guayaquil* was transferred from 2012-10-09 to 2012-10-12, which means it was celebrated on 2012-10-12. Days that are type *Bridge* are extra days that are added to a holiday (e.g., to extend the break across a long weekend). These are frequently made up by the type *Work Day* which is a day not normally scheduled for work (e.g., Saturday) that is meant to payback the Bridge.
- *Additional* holidays are days added a regular calendar holiday, for example, as typically happens around Christmas (making Christmas Eve a holiday).

4.9 Additional Notes

- Wages in the public sector are paid every two weeks on the 15 th and on the last day of the month. Supermarket sales could be affected by this.
- A magnitude 7.8 earthquake struck Ecuador on April 16, 2016. People rallied in relief efforts donating water and other first need products which greatly affected supermarket sales for several weeks after the earthquake.

5 Competition Rules

https://www.kaggle.com/c/favorita-grocery-sales-forecasting/rules

- Start Date: October 19, 2017
- Entry Deadline: January 8, 2018
- End Date: January 15, 2018 @ 11:59 PM UTC
- COMPETITION TITLE: Corporación Favorita Grocery Sales Forecasting
- COMPETITION SPONSOR: Corporación Favorita C.A.
- COMPETITION WEBSITE: https://www.kaggle.com/c/favorita-grocery-sales-forecasting

- First Prize: $15,000
- Second Prize: $10,000
- Third Prize: $5,000
- EXTERNAL DATA: Permitted, with guidelines as follows, Use of publicly available external data is permitted, provided it does not pertain to Corporación Favorita nor to any entity in the same line of business as Corporación Favorita. For example, you may use pre-trained networks or natural language tools/libraries. The source of any external data must be posted to the official competition forum prior to the First Submission Deadline.
- Non-Exclusive License: Outside of Ecuador, you will grant to Competition Sponsor and its designees a worldwide, non-exclusive, sub-licensable, transferable, fully paid-up, royalty-free, perpetual, irrevocable right to use, reproduce, distribute, create derivative works of, publicly perform, publicly display, digitally perform, make, have made, sell, offer for sale and import your winning Submission and the source code used to generate the Submission, in any media now known or hereafter developed, for any purpose whatsoever, commercial or otherwise, without further approval by or payment to Participant.

6 Descripción del Trabajo a Realizar

6.1 Carga y Preparación de los Datos

6.2 Análisis Exploratorio de Datos

- Visualización datos train
- ventas en general,
- por tienda,
- por tipo de tienda,
- por geolocalización,
- por estado,
- por ciudad,
- ventas en el tiempo,
- correlación ventas – precio del petróleo,
- días de fiesta por estado
- correlación ventas – días de fiesta,
- análisis de los productos y sus familias,
- Estudio de ventas de los artículos en promoción
- frecuencia de venta de productos
- clasificación perecederos – no perecederos
- estudio de las transacciones en el tiempo,

6.3 Estudio Predictivo

para lo que utilizaré los algoritmos con los que obtenga un resultado que me permitan obtener una buena clasificación

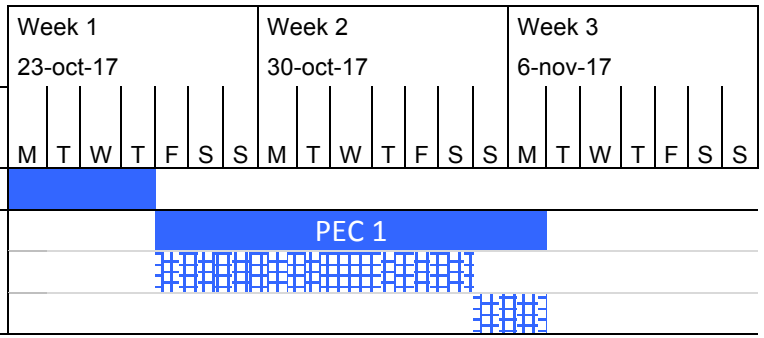
6.4 Planificación

B2.342 Trabajo Final de Master - AD

Corporación Favorita Grocery Sales Forecasting - Kaggle

Author: Gabriel Kreplak

WBS	Task	Start	End	Cal. Days
1	Theme Selection	22/10/17	26/10/17	5
2	Initial Proposal	27/10/17	6/11/17	11
2.1	Preliminary Analysis	27/10/17	4/11/17	9
2.2	Preliminary design	5/11/17	6/11/17	2

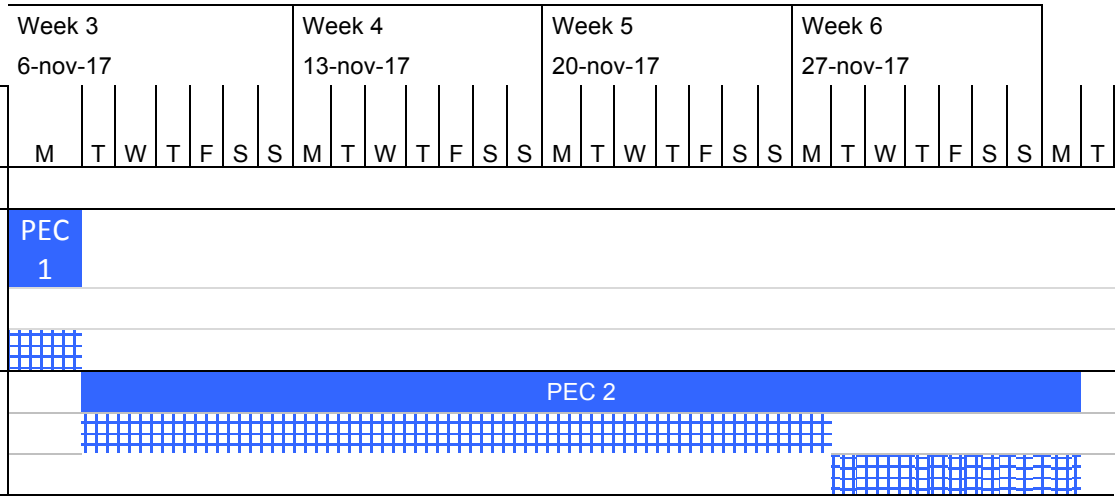


B2.342 Trabajo Final de Master - AD

Corporación Favorita Grocery Sales Forecasting - Kaggle

Author: Gabriel Kreplak

WBS	Task	Start	End	Cal. Days
1	Theme Selection	22/10/17	26/10/17	5
2	Initial Proposal	27/10/17	6/11/17	11
2.1	Preliminary Analysis	27/10/17	4/11/17	9
2.2	Preliminary design	5/11/17	6/11/17	2
3	Memory Structure	7/11/17	4/12/17	28
3.1	EDA design	7/11/17	28/11/17	22
3.2	Preliminary Predict	29/11/17	4/12/17	6

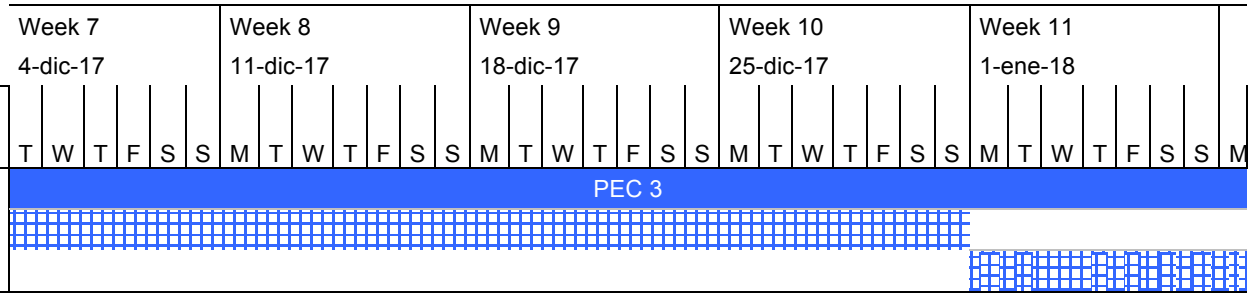


B2.342 Trabajo Final de Master - AD

Corporación Favorita Grocery Sales Forecasting - Kaggle

Author: Gabriel Kreplak

WBS	Task	Start	End	Cal. Days
4	First Version	5/12/17	8/1/18	35
4.1	Prediction Tuning	5/12/17	31/12/17	27
4.2	Report Issuing	1/1/18	8/1/18	8

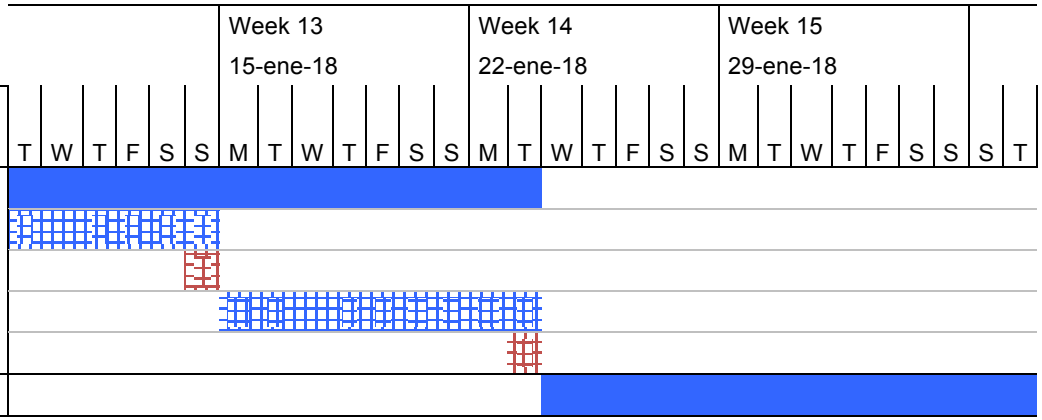


B2.342 Trabajo Final de Master - AD

Corporación Favorita Grocery Sales Forecasting - Kaggle

Author: Gabriel Kreplak

WBS	Task	Start	End	Cal. Days
5	Delivery	9/1/18	23/1/18	15
5.1	Last review	9/1/18	14/1/18	6
	Milestone: Submittal	14/1/18	14/1/18	
5.2	Report Correction	15/1/18	23/1/18	9
	Milestone: Delivery	23/1/18	23/1/18	
6	Evaluation	24/1/18	7/2/18	15



7 Resto de capítulos

En estos capítulos, hay que describir los aspectos más relevante del diseño y desarrollo del proyecto, así como de los productos obtenidos. **La estructuración de los capítulos puede variar según el tipo de Trabajo**.

En cada apartado es muy importante describir las alternativas posibles, los criterios utilizados para tomar decisiones y la decisión tomada.

En caso de que corresponda, se incluirá un apartado de “Valoración económica del trabajo”. Este apartado indicará los gastos asociados al desarrollo y mantenimiento del trabajo, así como los beneficios económicos obtenidos. Hacer un análisis final sobre la viabilidad del producto.

8 Conclusiones

Este capítulo tiene que incluir:

- Una descripción de las conclusiones del trabajo: Qué lecciones se han aprendido del trabajo?.
- Una reflexión crítica sobre el logro de los objetivos planteados inicialmente: Hemos logrado todos los objetivos? Si la respuesta es negativa, por qué motivo?
- Un análisis crítico del seguimiento de la planificación y metodología a lo largo del producto: Se ha seguido la planificación? La metodología prevista ha sido la adecuada? Ha habido que introducir cambios para garantizar el éxito del trabajo? Por qué?
- Las líneas de trabajo futuro que no se han podido explorar en este trabajo y han quedado pendientes.

9 Glosario

Definición de los términos y acrónimos más relevantes utilizados dentro de la Memoria.

10 Bibliografía

Lista numerada de las referencias bibliográficas utilizadas dentro de la memoria. En cada lugar donde se utilice una referencia dentro del texto, hay que indicarla citando el número de la referencia, por ejemplo: [7].

Es muy importante incluir **todas** las referencias utilizadas y citarlas apropiadamente, es decir, incluyendo toda la información necesaria para identificar la referencia. La información mínima que hay que incluir según el tipo de referencia es:

- **Libro:** Autores, Título, Edición (si se tercia) Editorial, Ciudad, Año.
- **Artículo de revista:** Autores, Título, Nombre de la Revista, Número de Página inicial y final, Número de la revista / Volumen, Año.
- **Web:** URL y fecha en que se ha visitado.

11 Anexos

Listado de apartados que son demasiado extensos para incluir dentro de la memoria y tienen un carácter autocontenido (por ejemplo, manuales de usuario, manuales de instalación, etc.)

Dependiente del tipo de trabajo, es posible que no haya que añadir ningún anexo.