Retail Sales Forecasting: Corporacion Favorita

Performing Better Business Decisions

https://github.com/amjadraza/retail-sales-prediction

Who am I and What am I doing?

Muhammad A. Raza (Ph.D.)

Data Scientist
Date Engineer

Muhammad Amjad Raza is an Electrical Engineer and Data Scientist with more than 10 years of professional experience in various industries. He is working with The Portland House group as quant for last three years.

Retail Sales Forecasting

- Corporacion Favorita data
- Model to forecast
- Why Business need Model

Machine Learning Approach

- Why we need Model
- How do I approach it?
- Proposed Model
- Future Improvements
- Tools/Technologies Used

A production ready, Scalable retail sales forecasting machine learning pipeline

Corporacion Favorita & Julian Assange

Corporacion Favorita-

TOTAL ASSETS

$1,440,2 Millions

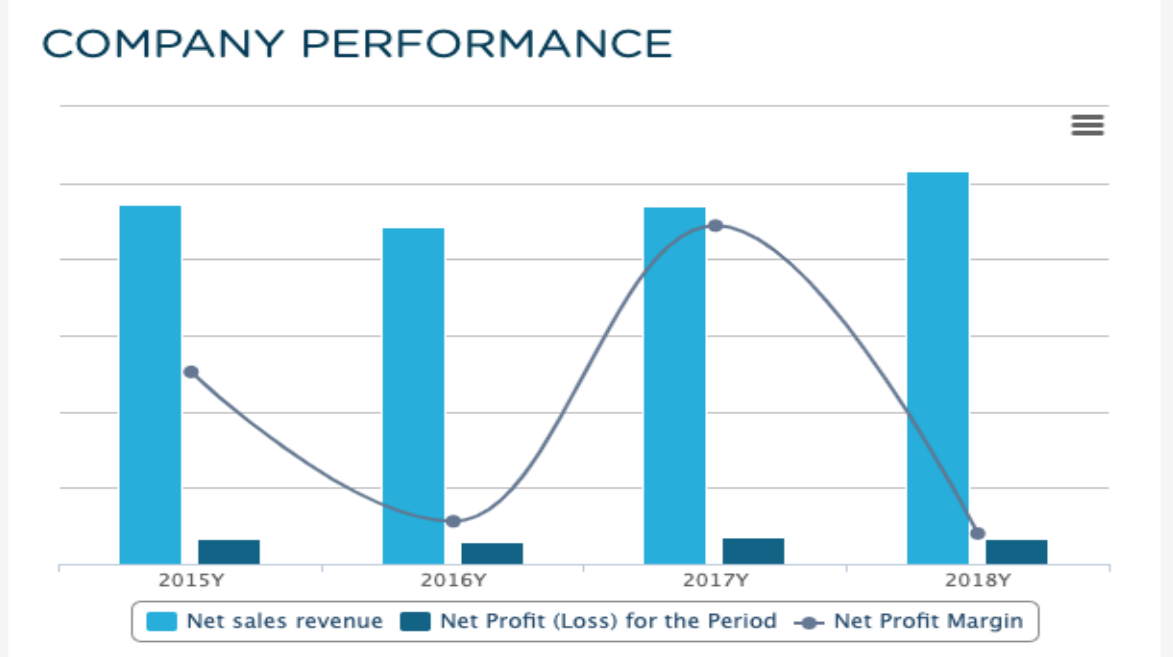
Revenue

$2045 Millions

- large Ecuadorian-based grocery retailer
- Public Listed Company
- Founded in 1945
- A group of 20+ companies

Net Income

$154 Millions



How Sales Forecasting helps in Business?

01. Logistics Planning

- Optimal logistic transportation
- Optimal storage capacity
- Staff rostering

02. Better Business Forecast

- Understand future revenue
- Helps to find demand and supply
- Better revenue predictions over short periods as well as long period

03. Promo optimization

- Store sales planning
- Hot promotions planning

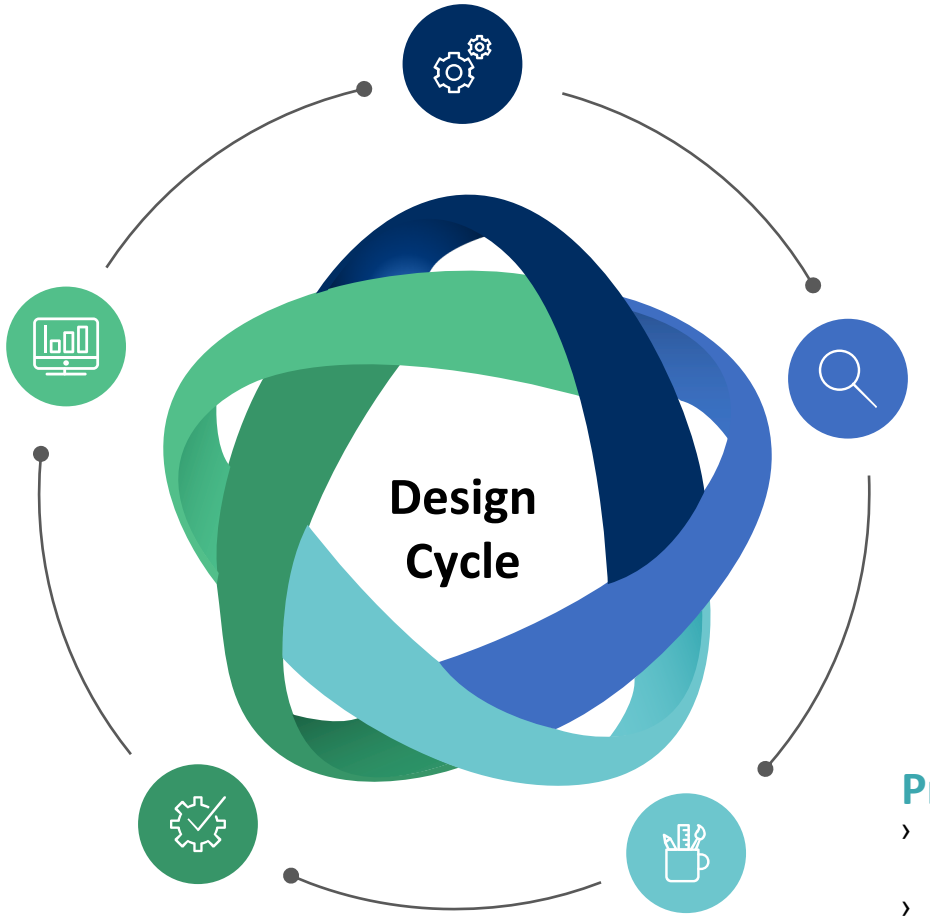
04. Consumer behavior

- What local consumer like the most
- Personalized consumer demand
- Meet individual consumer demand even before they ask

An Overview of Forecasting ML Framework

Exploratory Data Analysis

- Data Engineering & Governess
- Distributions of data
- Can we believe the data?



Feature Engineering

- Relationship between data tables
- Extracting the features from existing data
- Creating new features using domain specific knowledge

Preparing Data for Model

- Preparing the training, validation and test data
- Imputing missing values if necessary with some logic

Model Training

- Train the model
- Find optimal/best model
- Save the model in pickle

Deploy the Model for Production

- Deploy the Model on Cloud or in-house for production.
- Keep tracking of real time performance of model

Exploratory Data Analysis

1. Store Data Table
2. Items Data Table
3. Transactions Data Table
4. Holidays Data Table
5. Oil Data Table

EDA-Part 1/1

EDA-Part 2/2

Feature Engineering

Understanding the Training Data

1. Size of training data
   - Billions of rows of data
   - Cant handle on simple home computers
2. Identify Target variable
   - Target Variable is the unit sales for each store and each item
   - Frequency of predicting could be daily, weekly, monthly.
3. Problem Identified
   - Missing data points filling.
   - Original solution is filling with zero.
   - forward filling average of last 20 days unit sale average

Preparing the Data For Training

1. Training data set
   - Using the winning solution to build upon
   - First half of 2017 for training
2. Validation set
   - Validation set is used to choose the best model
   - Month of validation period
3. Test Set
   - Test set is given starting from Aug-2017 to the end

Feature Engineering notebook

Machine Learning Model

Light GBM

1 What I chose Light GBM

- Gradient based framework written for faster training
- Grows trees vertical

2 Industry Use Cases

Light GBM is very famous among Kaggle community
Industry loves its faster training time

3 Fast and Scalable

- Support for GPU Training
- Support for Spark training using MS library

Model Comparison

	Max Round	Num_leaves	Metric	Validation MSE	Weight MSE
Model-Winner	200	3	l2	0.392	**0.6266**
Model-Proposed_v1	200	3	l2	0.2206	**0.4699**
Model-Proposed_v2					

Comments

The training parameters are not optimized due to computation time

Need to compare the test error –out of sample

Further study need to make sure no time machine

Model Training notebook

Future work to complete the pipeline

01

Feature Engineering

One of the interesting task, I would like to complete is creating new features.

02

ML/DL Models

Adding options for more models. Use linear Bayesian model to combine the predictions from individual models

---

03

Distributed Computation

With the increase of data, we can scale this pipeline using spark capabilities. Light GBM already has support to run with pyspark.

04

Consistency check of Model

Designing a framework to monitor the consistency of Model(s). Rolling training and validation is one option.

05

Deploying with CI/CD

Deploying the model for production. Set up CI/CD and add the test coverage.

Data Science Tools/Technologies