**Kaggle Knowledge**

<https://www.linkedin.com/pulse/my-first-gold-medal-kaggle-competitions-antonio-marin/?trackingId=JZPGvcObBPRvdu3U1QxfwQ%3D%3D>

# Forecasting -M5

Competition Summary

<https://www.kaggle.com/c/m5-forecasting-accuracy/discussion/163414>

1. [https://www.kaggle.com/headsortails/back-to-predict-the-future-interactive-m5-eda (1700](https://www.kaggle.com/headsortails/back-to-predict-the-future-interactive-m5-eda%20(1700) updoots)
2. <https://otexts.com/fpp2/bootstrap.html>
3. <https://stats.stackexchange.com/questions/41145/simple-way-to-algorithmically-identify-a-spike-in-recorded-errors>
   1. If >2σ, then there’s a spike
   2. <https://stats.stackexchange.com/questions/64314/features-to-represent-peaks-in-time-series>
   3. <https://www.real-statistics.com/descriptive-statistics/symmetry-skewness-kurtosis/#:~:text=Kurtosis%20pertains%20to%20the%20extremities,)%20%3D%20the%20kurtosis%20of%20S>.
   4. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2631518/> (peak detection)
      1. <https://stackoverflow.com/questions/1713335/peak-finding-algorithm-for-python-scipy>
   5. <https://stackoverflow.com/questions/26632205/finding-a-spike-or-drop-in-a-dataset-programatically>
   6. Chart, waterfall chart

      Description automatically generated
   7. [tweedie loss function for right skewed data with long tail](https://towardsdatascience.com/tweedie-loss-function-for-right-skewed-data-2c5ca470678f)
   8. [DeepAR Forecasting Model](https://docs.aws.amazon.com/sagemaker/latest/dg/deepar.html)
   9. <https://ts.gluon.ai/>

WRMSSE vs MASE vs RMSE

1. <https://www.kaggle.com/c/m5-forecasting-accuracy/discussion/134044>

Best competition for beginners

<https://www.kaggle.com/getting-started/44088>

Getting started with deep learning

1. <https://www.kaggle.com/andradaolteanu/how-i-taught-myself-deep-learning-vanilla-nns>

50 advanced data science tips for tabular data

<https://www.kaggle.com/vbmokin/50-advanced-tips-data-science-for-tabular-data>

Categories

1. Computer Vision
2. NLP
3. Time series modelling
4. EDA
5. Data Visualization
6. Users to follow

Solutions

1. <https://www.kaggle.com/c/shopee-product-matching/discussion/225369>
2. <https://www.kaggle.com/sudalairajkumar/winning-solutions-of-kaggle-competitions>
3. <https://www.kaggle.com/c/nfl-impact-detection/discussion/208979> (2nd place solution)
4. <https://www.kaggle.com/theoviel/2nd-place-ncaaw-2021> (2nd place solution) (logistic regression)
5. Look for vinbig x ray solution – that is part of object detection

Object detection

1. <https://www.kaggle.com/brsdincer/vehicle-detection-image-set> (vehicle detection)

deep learning research papers

<https://www.kaggle.com/c/shopee-product-matching/discussion/224583>

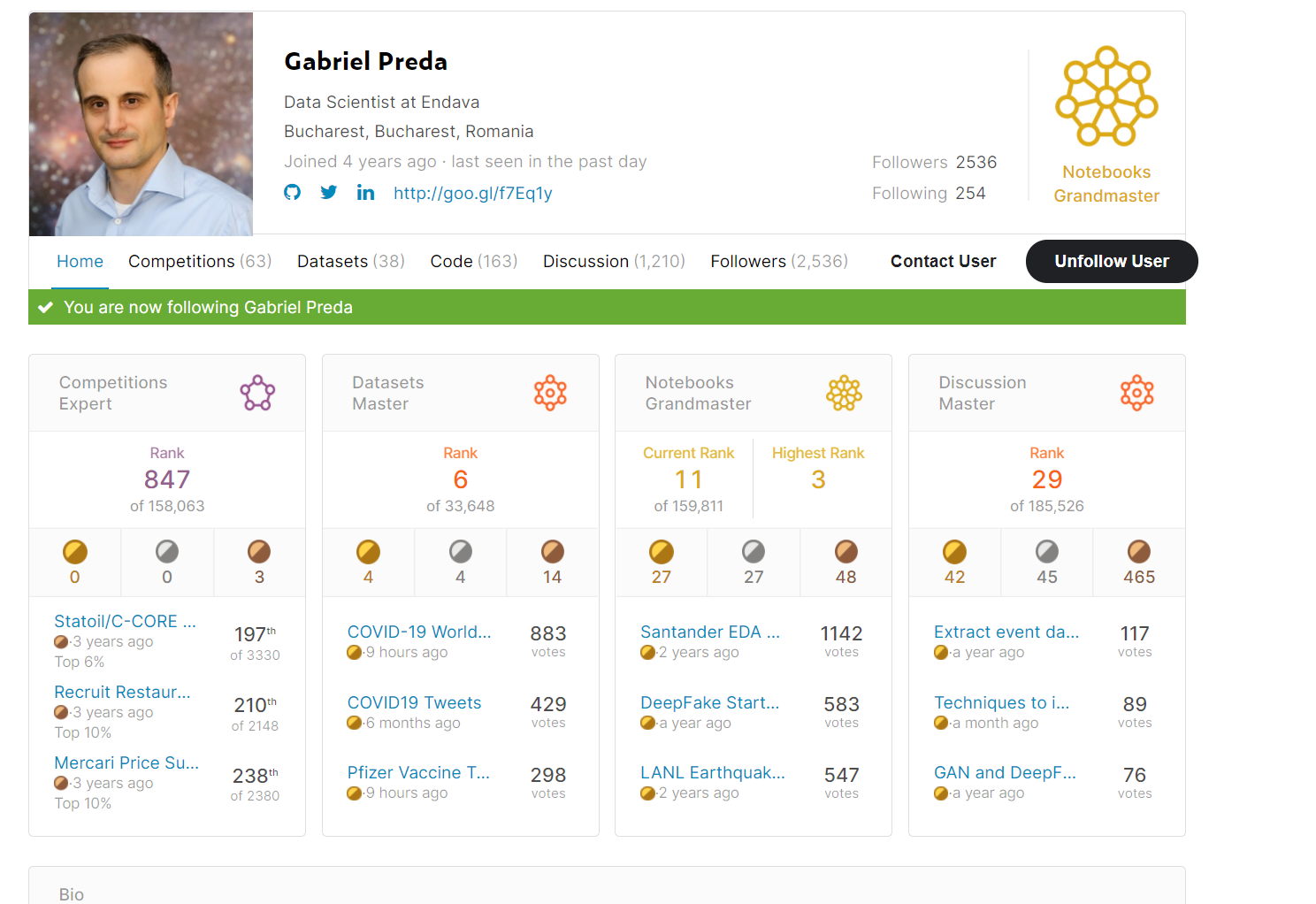
GEOSPATIAL

1. <https://www.kaggle.com/alexisbcook/your-first-map>

ALL IN ONE KAGGLE

<https://www.kaggle.com/ldfreeman3/a-data-science-framework-to-achieve-99-accuracy> (good source)

1. <https://github.com/zwcolin/dsc190-kaggle>
2. Mark peng



# Good starting point to start exploring Kaggle

1. <https://www.kaggle.com/c/tabular-playground-series-feb-2021/overview>

# ECG

1. <https://www.nature.com/articles/s41597-020-0495-6>
2. <https://www.kaggle.com/khyeh0719/ptb-xl-dataset>
3. <https://www.kaggle.com/khyeh0719/ptb-xl-dataset-wrangling>

Kaggle users

1. <https://www.kaggle.com/rankings>
2. <https://www.kaggle.com/subinium/the-hitchhiker-s-guide-to-the-kaggle/comments>
   1. Check out the authors web page SUBIN AN <https://subinium.github.io/how-to-become-a-kaggle-kernel-master/>

EDA (all types, didn’t rank good to bad) Commerce

1. <https://www.kaggle.com/ekami66/detailed-exploratory-data-analysis-with-python>
2. <https://www.kaggle.com/khyeh0719/ptb-xl-dataset-wrangling>
3. <https://www.kaggle.com/gpreda/2019-data-science-bowl-eda>
4. <https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python>
5. <https://towardsdatascience.com/a-gentle-introduction-to-exploratory-data-analysis-f11d843b8184>

https://www.kaggle.com/pmarcelino/comprehensive-data-exploration-with-python

https://towardsdatascience.com/a-gentle-introduction-to-exploratory-data-analysis-f11d843b8184

research questions example

What was the best month for sales? How much was earned that month?

Which state had the highest number of sales

what time should we display advertisement to maximized likehood of customer buying product

Which products are most often sold together

Which product sold the most? Why do you think it sold the most?

Retail

1. <https://www.kaggle.com/manjeetsingh/retaildataset/code?datasetId=2296&sortBy=voteCount>
2. Packaging for pop products
3. Volume – 1.3L to 1.0L same price, higher margin, bottle size has become bigger
4. Acive ingredients – sugar go up, coco go down
5. FMCG, NPD,
6. High freq of out of stock 🡪 forecasting issue , key accounts, jaya grocer, vollage grocer, 99speed mart etc
7. Size of the product and warehouse size
8. Smaller stores 🡪 product close to cashier tends to sell more
9. Confidence interval
10. Anomaly
11. Dashboards
    1. What is actionalble on one side
    2. What is non actionalble on the other side
    3. One section needs to be very clear actionables
    4. For the last 30 days, how frequent are your stores going out of stock
12. Product level
    1. What product are currently out of stock
    2. Problem is long tail, oh I can live with that (long tail concept)
13. Three markets
    1. Hot zone
    2. Cashiers
    3. Back of the aisle
14. Coverage and weighted distributions and go by state
15. Inventory vs demand
16. Value added data analytics

Dataset

<https://www.kaggle.com/manjeetsingh/retaildataset/code?datasetId=2296&sortBy=voteCount>

* <https://www.kaggle.com/aremoto/retail-sales-forecast>
* <https://www.kaggle.com/shubhamsinghgharsele/retail-data-analysis>

how models work

1. <https://www.kaggle.com/dansbecker/how-models-work>
2. <https://www.kaggle.com/artgor/code> - quite bit of EDA

kyle peng Malaysia

1. <https://medium.com/kaggle-blog/dstl-satellite-imagery-competition-1st-place-winners-interview-kyle-lee-6571ce640253>

Simple Matplotlib & Visualization Tips

1. <https://www.kaggle.com/subinium/simple-matplotlib-visualization-tips>

# EDA

1. <https://www.kaggle.com/niharika41298/netflix-visualizations-recommendation-eda>
2. <https://www.kaggle.com/shivamb/netflix-shows-and-movies-exploratory-analysis>
3. <https://www.kaggle.com/nitindatta/students-performance-in-exams-eda-in-depth>
4. <https://www.kaggle.com/allunia/m5-sales-uncertainty-prediction> (good, check first part of viz)( <https://otexts.com/fpp2/bottom-up.html>)
5. <https://www.kaggle.com/dimitreoliveira/model-stacking-feature-engineering-and-eda>
6. <https://www.kaggle.com/yasufuminakama/inchi-resnet-lstm-with-attention-starter>
   1. Convert image to text

E-commerce

<https://www.kaggle.com/olistbr/brazilian-ecommerce>

<https://www.kaggle.com/thiagopanini/e-commerce-sentiment-analysis-eda-viz-nlp>

<https://www.kaggle.com/duygut/brazilian-e-commerce-data-analysis>

<https://www.kaggle.com/jsaguiar/e-commerce-exploratory-analysis>

<https://www.kaggle.com/kabure/simple-eda-sales-and-customer-patterns>

Predicting future sales data

<https://www.kaggle.com/c/competitive-data-science-predict-future-sales/discussion/74835>

<https://otexts.com/fpp2/hts.html>

Categorical EDA

https://www.kaggle.com/subinium/tps-may-categorical-eda

Forecasting

1. M5
2. <https://otexts.com/fpp2/hts.html>

ML

Easy

<https://www.kaggle.com/vishalyo990/prediction-of-quality-of-wine>

**1. Random Forest Classifier**

**2. Stochastic Gradient Descent Classifier**

**3. Support Vector Classifier(SVC)**

Then I use cross validation evaluation technique to optimize the model performance.

**1. Grid Search CV**

**2. Cross Validation Score**

Geospatail good maps

<https://www.kaggle.com/andresionek/geospatial-analysis-of-brazilian-e-commerce>

Kaggle notebook to walk through first submission

<https://www.kaggle.com/c/tabular-playground-series-feb-2021/overview>

# NLP

1. <https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/discussion/87868>
2. [Getting Started with NLP](https://www.kaggle.com/abhishek/approaching-almost-any-nlp-problem-on-kaggle)
3. <https://www.kaggle.com/ruchi798/commonlit-readability-prize-eda-baseline>
   1. Good NLP EDA
4. <https://www.kaggle.com/andradaolteanu/i-commonlit-explore-xgbrf-repeatedfold-model>
   1. Good explanation when data labelers don’t agree with each other

Shopee

<https://www.kaggle.com/c/shopee-product-matching/code?competitionId=24286&sortBy=voteCount>

computer vision

<https://www.kaggle.com/c/cdiscount-image-classification-challenge/discussion/45863> (very good model explanation) (with hardware specifications) (<https://drive.google.com/drive/folders/0B_DICebvRE-kb2dFd2FKX1hfRkE>)

House prices

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

Underrated data science notebook

1. <https://www.kaggle.com/headsortails/hidden-gems-a-collection-of-underrated-notebooks>

Rank 1 Kaggler

1. <https://www.kaggle.com/cdeotte/25-million-images-0-99757-mnist>

# Kaggle Project planning

A template for every kind of dataset

1. Which industry?
2. Feature engineering techniques?
   1. Image/video
   2. Sound
   3. Tabular

EDA notebooks

1. [M5-Forecasting-EDA](https://www.kaggle.com/headsortails/back-to-predict-the-future-interactive-m5-eda)

**The data:** We are working with **42,840 hierarchical time series**. [The data](https://www.kaggle.com/c/m5-forecasting-accuracy/data) were obtained in the 3 US states of California (CA), Texas (TX), and Wisconsin (WI). “Hierarchical” here means that data can be aggregated on different levels: item level, department level, product category level, and state level. The sales information reaches back from Jan 2011 to June 2016. In addition to the sales numbers, we are also given corresponding data on prices, promotions, and holidays. Note, that we have been warned that **most of the time series contain zero values**

1. To be worked on… (ctrf-F “EDA”)

# Timeseries

1. <https://www.kaggle.com/thebrownviking20/everything-you-can-do-with-a-time-series>

# Feature engineering

1. <https://www.kaggle.com/raddar/convert-dicom-to-np-array-the-correct-way>
2. <https://www.kaggle.com/c/hubmap-kidney-segmentation/discussion/232357>
3. <https://www.kaggle.com/gzuidhof/full-preprocessing-tutorial>
   1. Dicom

Model Stacking

1. <https://www.kaggle.com/arthurtok/introduction-to-ensembling-stacking-in-python>
2. Data professor youtube has a beginner’s tutorial on model stacking

# Project 1 – Convert Image to Text

<https://www.kaggle.com/c/bms-molecular-translation/data>

MNIST

* 1. Maybe we can use this to look for models
  2. <https://www.kaggle.com/c/digit-recognizer/code>

\*go to coursera and look at the MNIST notebook

To do list

1. Understanding the Data
2. How to Structure the data
3. How to ingest the data into the model
   1. Check MNIST model from the coursera course
4. How to train the data
5. How to output the results of the model

Project 2 – Feature Engineering Template