COMPETITIVE ML LEAGUE

An Introduction to Kaggle and Machine Learning

Moksh Jain

Student at National Institute of Technology Karnataka
Mumbai, Maharashtra, India
Joined 6 months ago · last seen 2 days ago
https://mj10.github.io

Competitions Contributor

Home | Competitions (4) | Discussion (1) | Contact User | Follow User

Competitions Contributor

Unranked

0	0	1

Mercari Price Suggestion C... · 16 days ago · Top 10%	222nd	of 2384
Webclub Recruitment · 6 months ago · Top 40%	4th	of 10
Toxic Comment Classificati... · 11 days to go · Top 23%	917th	of 4087

Kernels Contributor

Unranked

0	0	0

No kernel results

Discussion Contributor

Unranked

0	0	0

RNN_with_Keras_Ridge_SG... · a month ago	1	vote

Machine Learning

- Process of machines ‘learning’ to perform a specific task, using certain algorithms and/or statistical models.
- Similar to how humans learn, eg. Observing patterns, trying different strategies to reach a certain goal, being told what something means
- A subset of Artificial Intelligence, does not achieve all goals of AI.

Categories

- **Supervised Learning**: Given a set of inputs and corresponding outputs, learn the mapping from input to output. Eg. Humans learn in a similar way.
- **Unsupervised Learning**: Given some data, learn to discover intrinsic relations between the data, without any information regarding the relations. Eg. Given a set of colorful points, one can easily segregate them into groups based on colors even if he doesn't know what those colors.
- **Reinforcement Learning**: Given an environment which can be observed and provides rewards for actions taken, learn the best actions to be taken to maximize the reward. Eg. Learning to play games.

Supervised Learning

2 Major types of problems can be solved using Supervised Learning approach:

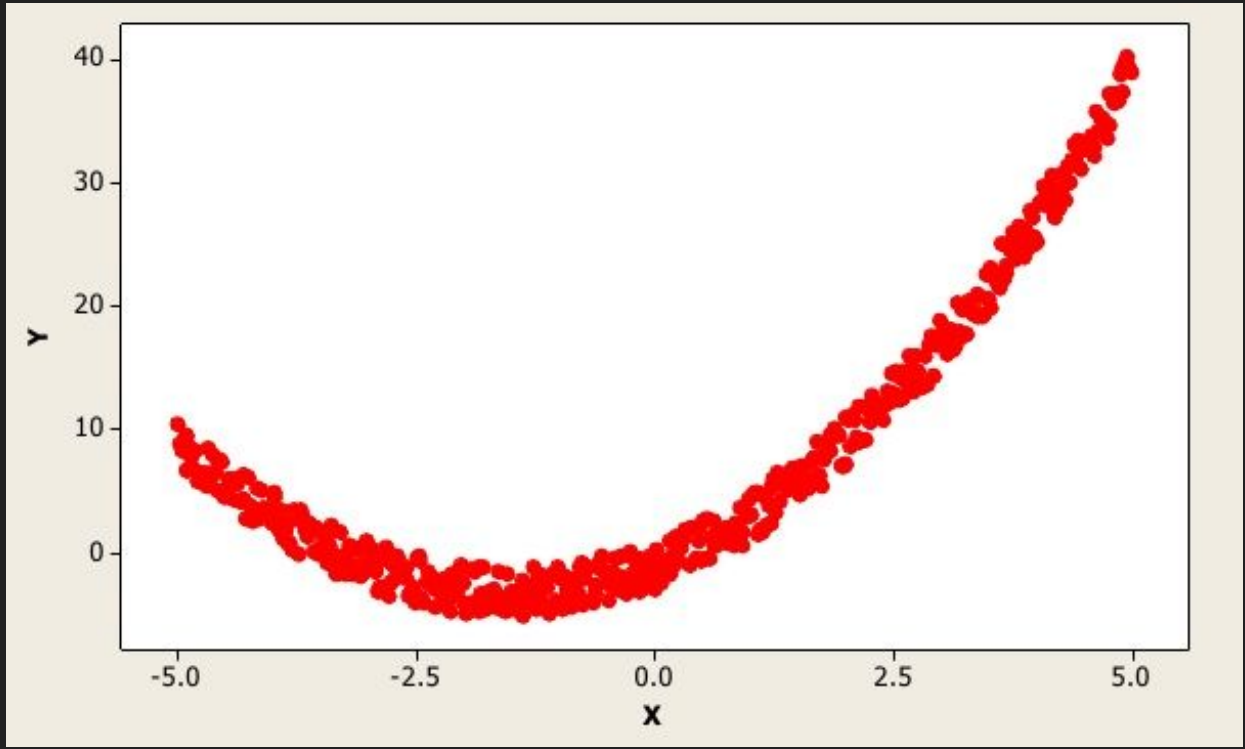
- **Classification**: You are given a set of images with corresponding discrete labels, learn to classify an unseen image and assign the proper label to it.
- **Regression**: You are given a set of inputs and corresponding outputs for a function, learn to predict the output of the function for any other input.

Essentially, the approach is to tell the model how accurate the prediction was and accordingly allow it to make a better prediction the next time.

The “Black Box”

- Machine Learning models can be understood as a black box, which takes some data as input and learns to predict the output, and can be used to predict output.
- A given problem can be solved using several different models, depending on the data available.
- *“A model is as good as the data it gets”*

Example



Example

- Assume you only have a model which can fit a line given the inputs and outputs.

$$y = a1 * x1 + a2 * x2 + \ldots + an * xn$$

- If you directly feed only the x values to the model, you might get incorrect results.
- One approach would be to introduce a new feature x2 = x ^ 2.
- Without changing the model, you were able to fit the data better, and achieve better results.
- Providing the model with the correct data, in the right form is very important.

Amlan Praharaj

Student at National Institute of Technology Karnataka
Karnataka, India
Joined a year ago · last seen in the past day

Followers 3
Following 8

Competitions Expert

Home | Competitions (15) | Kernels (1) | Discussion (4) | Datasets (1) | Followers (3) | Contact User | Follow User

Competitions Expert

Rank
632
of 78,207

0	3	3

Text Normalization Challen... · 4 months ago · Top 13%	33rd	of 260
Porto Seguro's Safe Driver ... · 3 months ago · Top 2%	72nd	of 5169
Mercari Price Suggestion C... · 16 days ago · Top 5%	98th	of 2384

Kernels Contributor

Unranked

0	0	0

Xgboost K-Fold 3 months ago	3	votes

Discussion Contributor

Unranked

0	0	0

Regarding any queries 5 months ago	2	votes
Regarding any queries 5 months ago	0	votes
Documented Code Submis... 5 months ago	0	votes

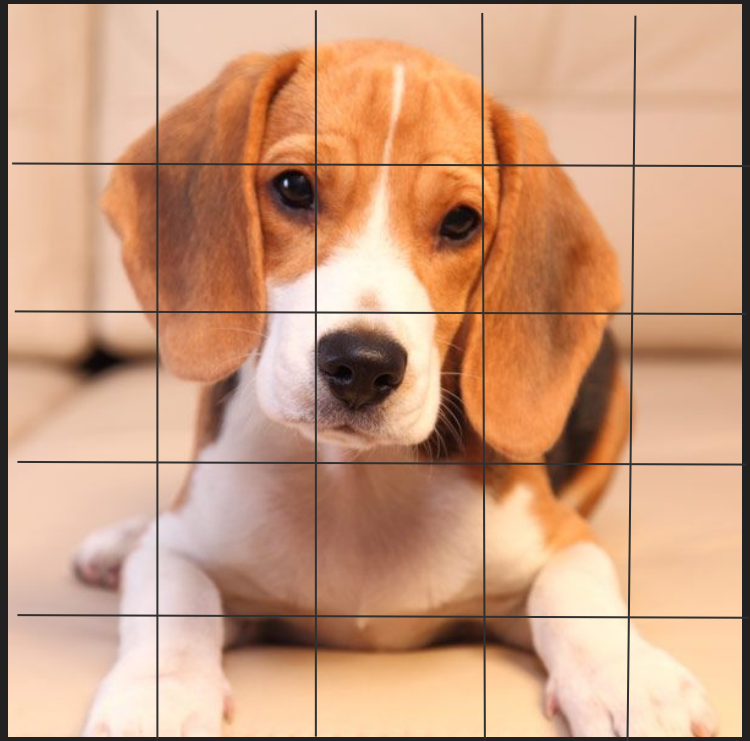
Features in images

What the black box sees

This image is composed of pixels, what is fed in as input to the model as a huge row of pixels, and its job is to find patterns that differentiate a dog from other pets such as cats.

But the performance of the model will be better if it could see in 2D. How do we do that. Let’s say we divide the image into grids and get numerical values from each grid. That lets us reduce the dimensionality of the image, and improves the input.

Vision in 2D



Forecasting the Future

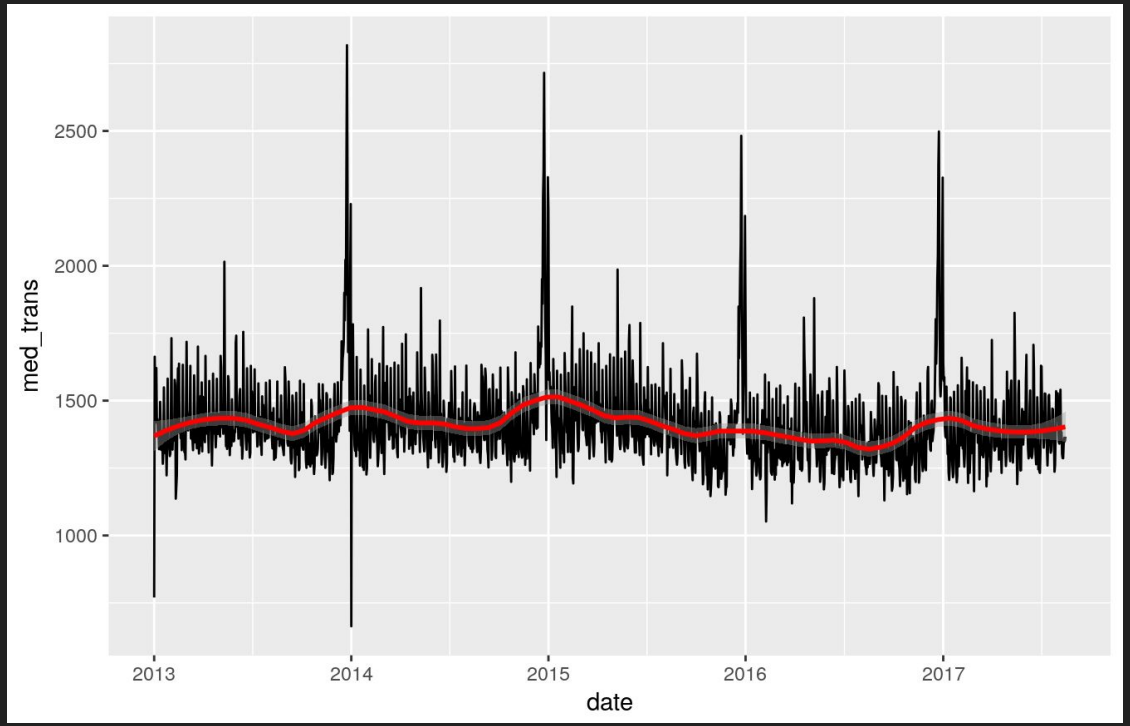
A very common machine learning task is using data to draw insights into the future.

It could involve predicting the prices of stocks, sales forecasting for inventory and predicting insurance claims among others.

We have a huge amount of prior knowledge that we use to predict the future, but how do we express the situational factors to the model?

One way to do this is time-series, i.e. keep track of past events using numerical columns.

A Time-series plot



Text

We can comprehend sentences due to our extensive language training, but the model can’t deal with words. The sentences have to be represented in a manner that the model can understand.

There are several ways of doing so.

Keeping track of the number of times a word occurs in a sentence could be a way to convert sentences to numerical columns.

Dealing with Tabular Data

All the problems discussed above are now Tabular Data problems that can be approached by any of the black box methods discussed by Moksh.

Future sessions will deal with choosing a suitable approach given the nature of the data.

somnath

Joined a year ago · last seen 23 days ago

Competitions Expert

Home | Competitions (3) | Contact User | Follow User

Competitions Expert

Current Rank	Highest Rank
2093 of 78,207	1751

0	1	1

Porto Seguro's Safe Driver ... · 3 months ago · Top 2%	72nd	of 5169
Text Normalization Challen... · 4 months ago · Top 34%	54th	of 162
Zillow Prize: Zillow's Home ... · 2 months ago · Top 51%	1,911th	of 3779

Kernels Novice

Unranked

0	0	0

No kernel results

Discussion Novice

Unranked

0	0	0

No discussion results

Porto Seguro

In this competition, we had to predict the probability that an auto insurance policy holder files a claim.

It was the largest Kaggle contest by participation.

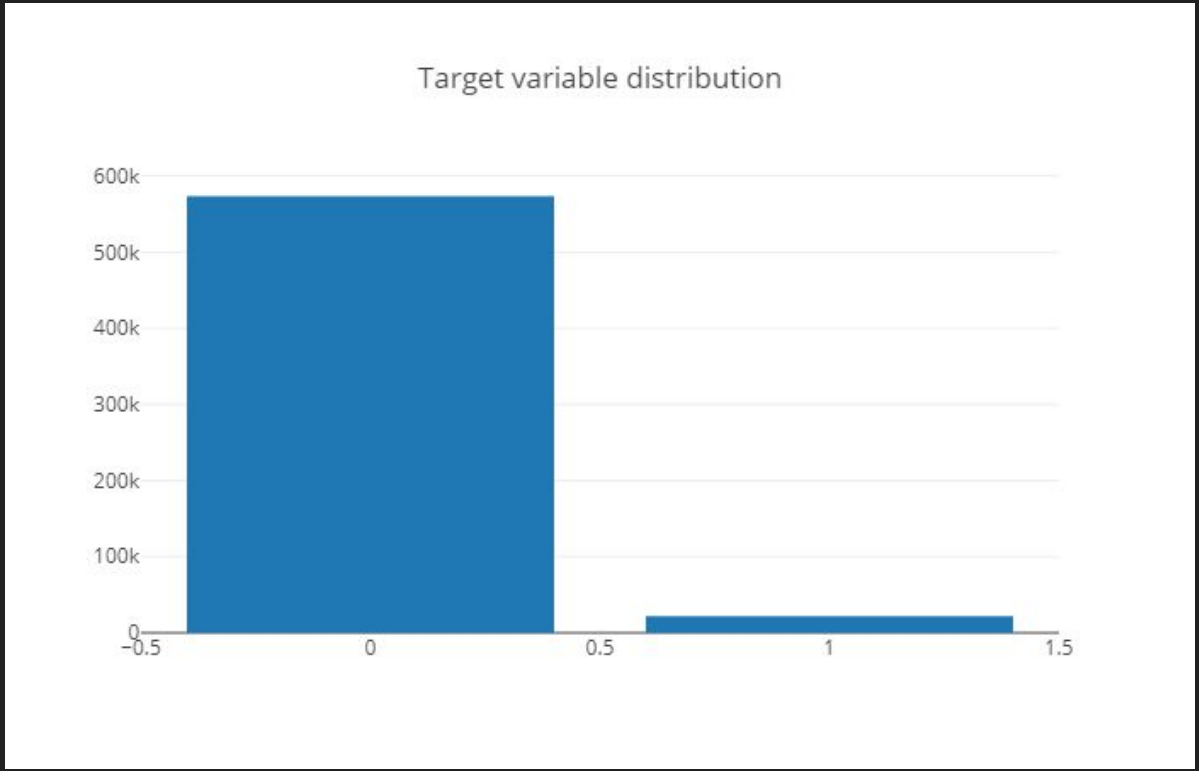
Sparse Dataset

Small Dataset

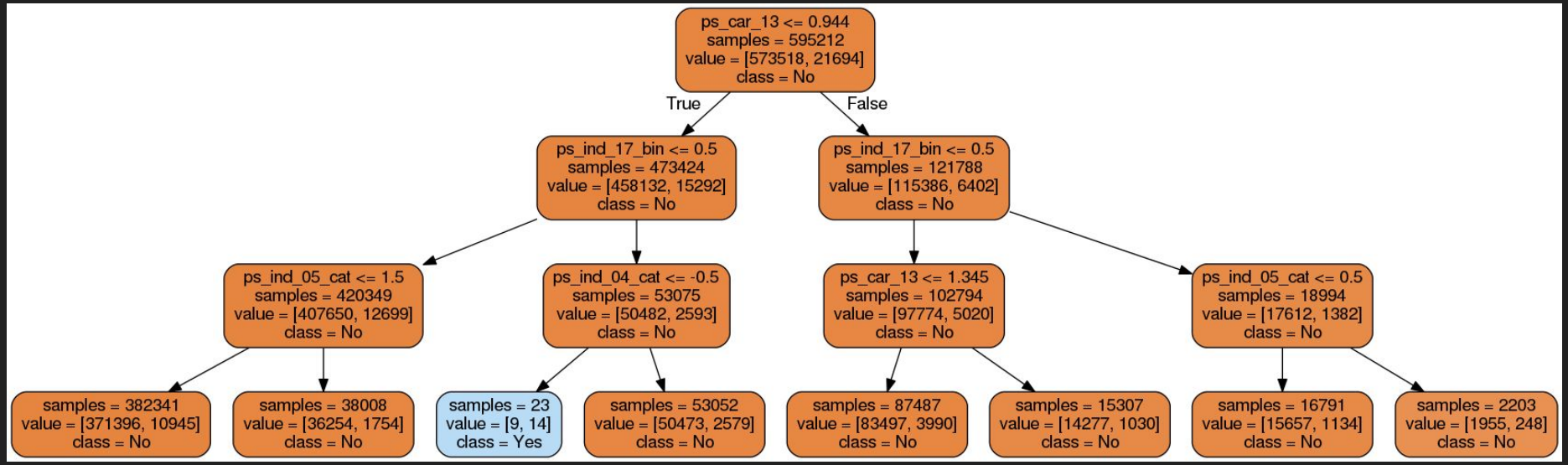
No feature engineering, complicated ensembles

Anonymised features

Porto Seguro



Porto Seguro



Porto Seguro

Field Aware Factorization Machines

Good for Imbalanced Data

Diverse Model for ensembles

Final Submission Involved a massive weighted ensemble of over 10 different models

Porto Seguro

Experiment: Diverse models, uncommon techniques, feature engineering, etc

An ensemble of weak predictors can give a strong result

Just because something should work “in theory”, does not mean the results would improve in practice

Simplify whenever possible: Prioritize simple ensembles, small models, etc.

Keep up to date with kernels and discussions and use their insights wherever possible

Ambareesh Prakash

Undergraduate Student at National Institute of Technology Karnataka
Mangalore, Karnataka, India
Joined 2 years ago · last seen in the past day

Followers 3
Following 7

Competitions Expert

Home | Competitions (14) | Kernels (1) | Discussion (8) | Followers (3) | Contact User | Follow User

Competitions Expert

Rank
581
of 78,207

0	4	1

Text Normalization Challen... · 4 months ago · Top 13%	33rd	of 260
Recruit Restaurant Visitor F... · a month ago · Top 4%	71st	of 2158
Porto Seguro's Safe Driver ... · 3 months ago · Top 2%	72nd	of 5169

Kernels Contributor

Unranked

0	0	0

No kernel results

Discussion Contributor

Unranked

0	2	3

My guess of the score high... · 4 months ago	6	votes
Looking for teammate(s) · 4 months ago	5	votes
Looking for teammate(s) · 4 months ago	2	votes

Future Sessions and Assignments

- Programming Language of Choice - Python
- Assignment - Setting up Python environment with necessary libraries(We recommend the Anaconda Python Distribution)
- Next Session - Build your first machine learning model on a dataset.
- Giving identities to the black boxes and taking them apart to see how they work.
- Getting to understand your data better.
- More on features and how to represent them so that your model can learn better.
- More on how to approach different kinds of problem statements with more advanced techniques.

A Sample of Different Models

- Linear Models
- Decision Trees
- Support Vector Machines
- Neural Networks
- Gradient Boosting
- Random Forests
- Factorisation Machines

Contact Information

- Mohit Reddy - mohitreddy1996@gmail.com
- Moksh Jain - mokshjn00@gmail.com
- Amlan Praharaj - amlan.prj@gmail.com
- Somnath Sarkar - somnath.k.sarkar@gmail.com
- Ambareesh Prakash - ambareesh.prakash@gmail.com