This project proposal is based on: https://www.kaggle.com/c/new-york-city-taxi-fare-prediction

Domain background.

In cities taxi is very popular. Help us to predict price for our customers, because They want to know possible price before ride.

Problem statement.

We are trying to predicting a fare amount for a taxi. We have given pickup and dropout location.

Datasets and inputs.

We use dataset provided by google.

- train.csv - Input features and target fare_amount values for the training set (about 55M rows).
- test.csv - Input features for the test set (about 10K rows). Your goal is to predict fare_amount for each row.
- sample_submission.csv - a sample submission file in the correct format (columns key and fare_amount). This file 'predicts' fare_amount to be $11.35 for all rows, which is the mean fare_amount from the training set.

Descripion of datafields:

ID:

- key - Unique *string* identifying each row in both the training and test sets

Features

- pickup_datetime- *timestamp* value indicating when the taxi ride started.
- pickup_longitude - *float* for longitude coordinate of where the taxi ride started.
- pickup_latitude - *float* for latitude coordinate of where the taxi ride started.
- dropoff_longitude - *float* for longitude coordinate of where the taxi ride ended.
- dropoff_latitude - *float* for latitude coordinate of where the taxi ride ended.
- passenger_count - *integer* indicating the number of passengers in the taxi ride.

Target

- fare_amount - `float` dollar amount of the cost of the taxi ride. This value is only in the training set; this is what you are predicting in the test set and it is required in your submission CSV.

Solution statement.

The solution will be the most accurate price for ride.

Benchmark model.

We can train model by using train dataset and test with test dataset. After training we measure difference between the predictions and the truth.

Evaluation metrics.

Evaluation metric for this project is the root mean-square error (RMSE). RMSE measusres difference between the predictions of a model and the ground truth.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y)^2}$$

where yi is the ith observation and $\hat{y}_i$ is the prediction for that observation.

Project design.

I dont know yet which architecture I should choose. I need to take these steps:

1. Deep analyze problem.
2. Data exploration.
3. Feature selection. Which is more important than other.
4. Select model. Maybe I should start with simple regression model. After that maybe I should choose more complex e.g. deep learning but with this size of dataset calculation takes a long time.
5. Train model, make predictions.
6. Benchmark.
7. Make decision if model is good enough. Train and test.