kaggle

NYC TAXI TRIP DURATION

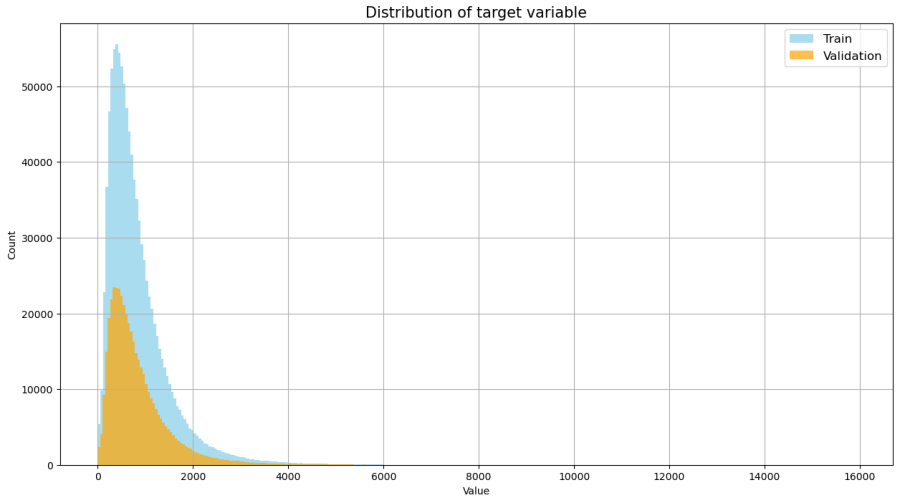
Christian Uccheddu - 800428
Federico Luzzi - 816753
Federico De Servi - 812166

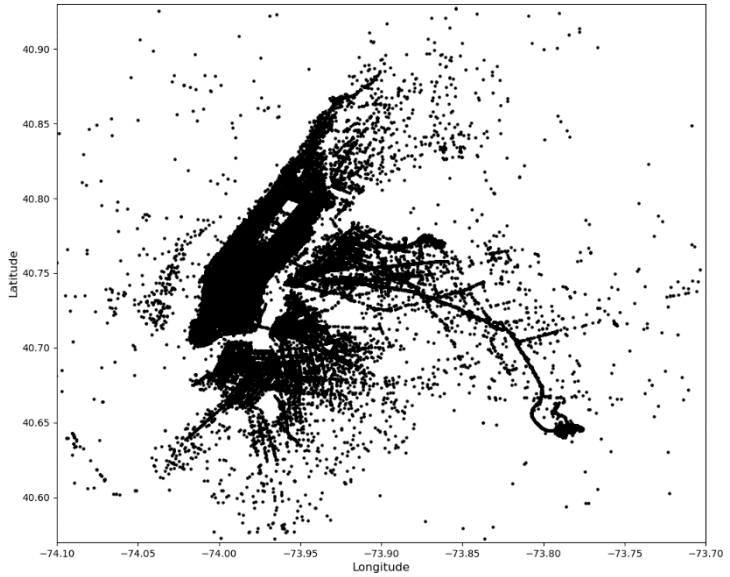
Image from https://www.cityandstateny.com/articles/opinion/editors-note/rough-stretch-nyc-taxi-limousine-commission.html

Image from https://fortune.com/2016/07/13/study-nyc-taxi-outperforming-uber/

Objective

Predict the taxi trip duration given the attributes in the dataset available on Kaggle





Original dataset

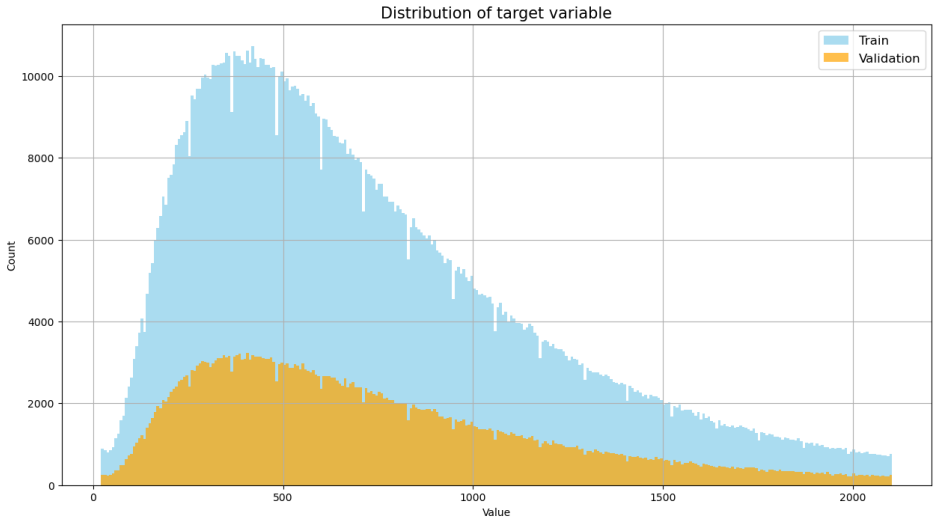
The dataset was taken from Kaggle and can be found at this link. The dataset presented some issues that had to be solved before applying our models:

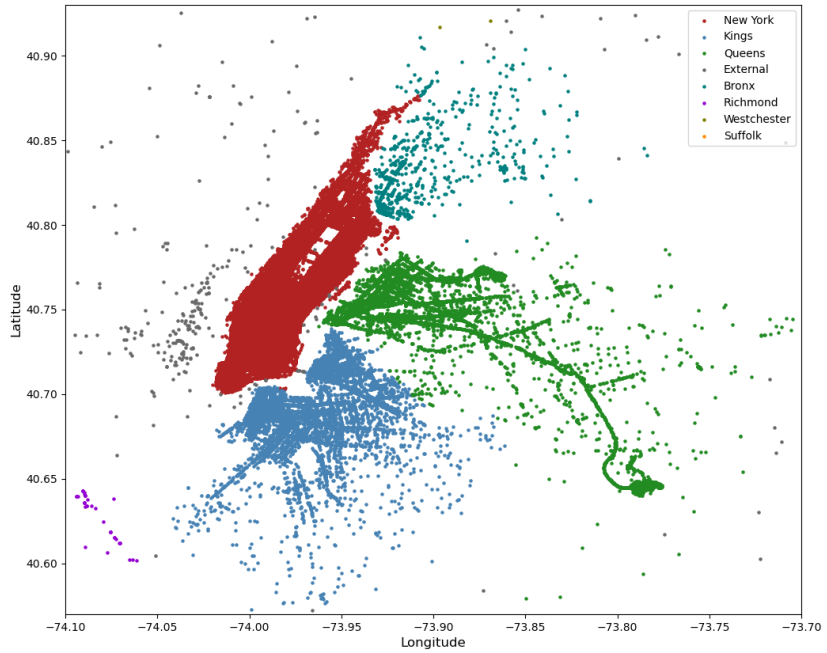
1) Some records contained presumably **wrong information** regarding the trip duration. (i.e. trips whose distance was 700 or more km completed in less than an hour)
2) Trips that presented a trip duration less than 10 seconds.
3) Geopositional data encoded in latitude and longitude. This Is not the ideal format to feed the models
4) Datetime attribute contained more than one information (day, time, weekday or n**ot**, etc)
5) No **distance** between the points is given. This had to be calculated.
6) No aggregate geographical info (county/block) was given.

Solutions

The solutions that had been applied were:

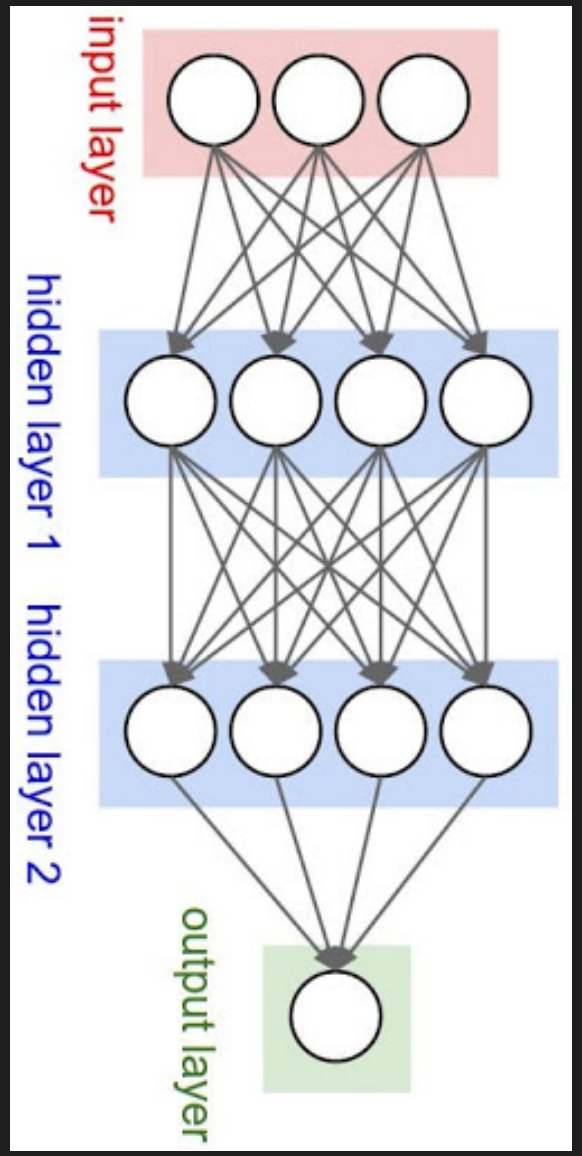
1) Records that contained presumably wrong information regarding the trip duration have been removed.
2) Trips that presented a trip duration less than 10 seconds have been removed.
3) Geopositional data encoded in latitude and longitude have been converted to x,y,z format.
4) Datetime attribute has been used to extract other attributes like hour, weekday or not.
5) The distance between points has been calculated using the Manhattan distance.
6) Geospatial information regarding Counties has been incorporated in our Kaggle dataset.





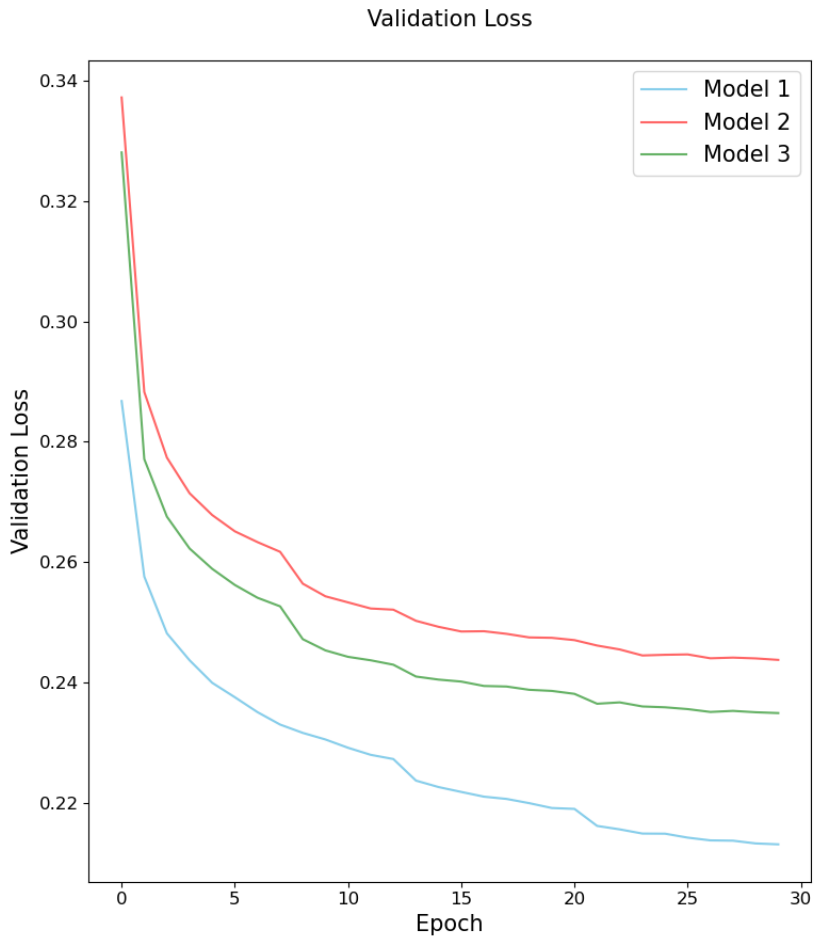
Methodological approach and evaluation

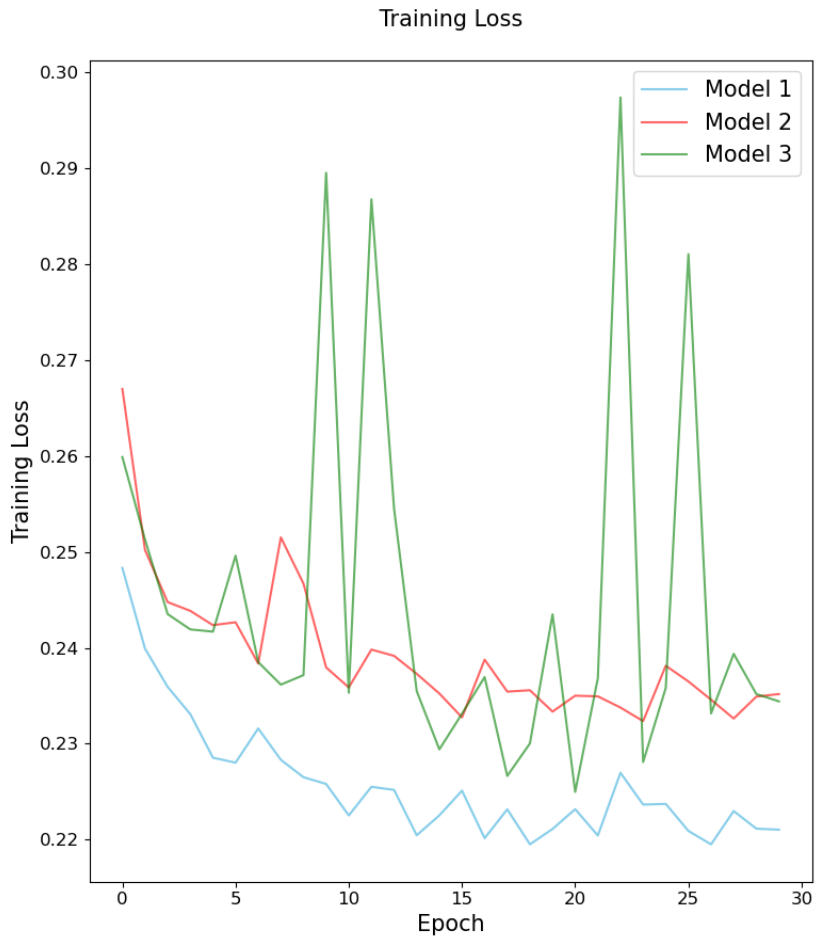
Given the large number of records, it was decided to use neural networks to solve the problem of predicting the taxi trip duration. The choice of the model and the hyperparameters was made after several empirical tests.



Comparison

The following graphs visualize the comparison between out top 3 models.





Best model: Model 1

This is model that resulted to be the best, in the end. It uses **Dense** layers, each with decreasing number of units and with the **ReLU** activation function, each followed by a **BatchNormalization** layer. This layer normalizes the results obtained from the previous dense layer, resulting in better and faster learning, and less overfitting.

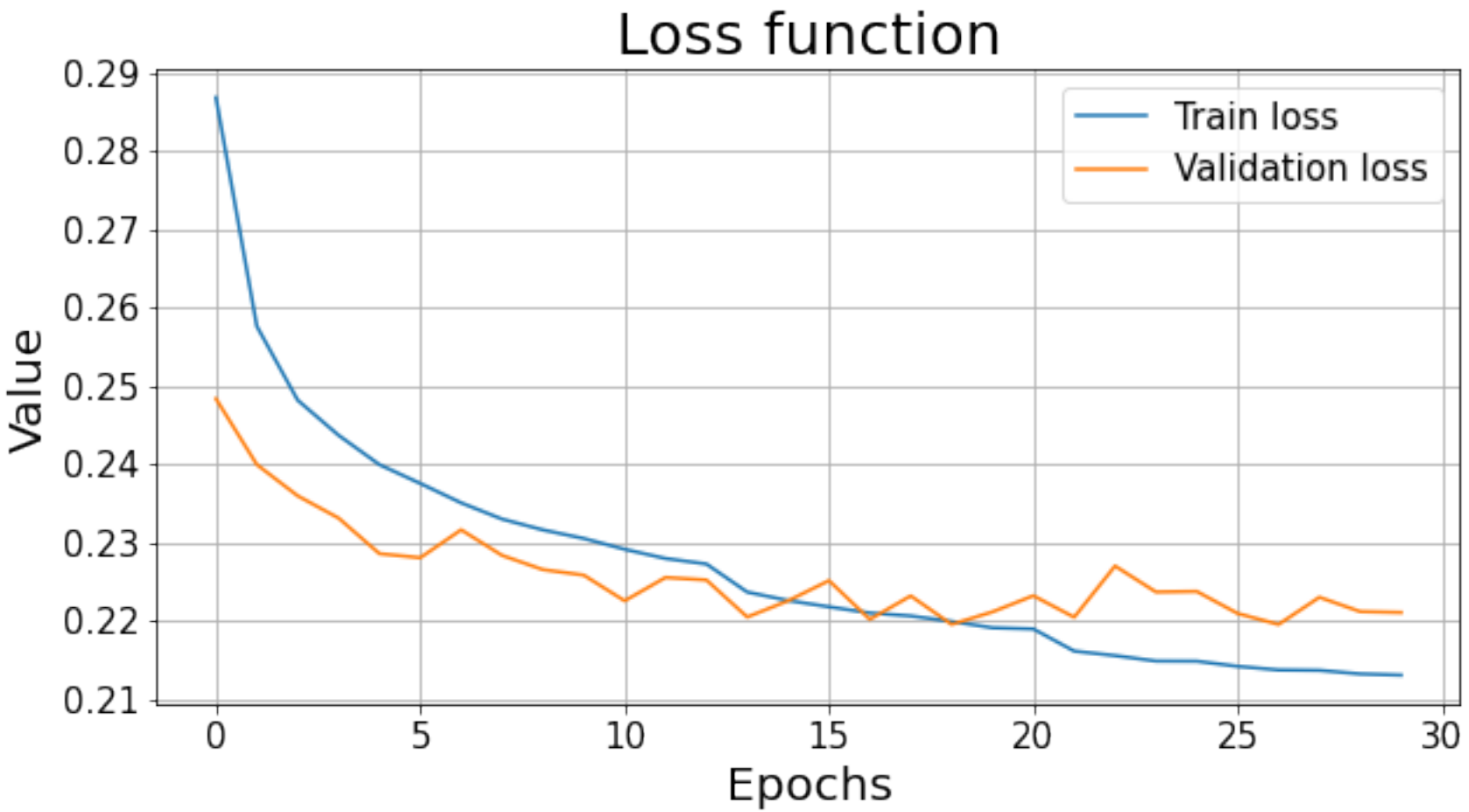
This layers all follows a **Gaussian Noise** layer, with very small hyperparameter. This is because adding noise to an underconstrained neural network model can have a regularizing effect and reduce overfitting. The **Learning Rate** changes going forward with the epochs in order to be able to quickly find an optimum point and then move carefully in the vicinity of this point so as not to skip it.

Layer (type)	Output Shape	Param #
gaussian_noise (GaussianNois	(None, 60)	0
dense (Dense)	(None, 512)	31232
batch_normalization (BatchNo	(None, 512)	2048
dense_1 (Dense)	(None, 256)	131328
batch_normalization_1 (Batch	(None, 256)	1024
dense_2 (Dense)	(None, 128)	32896
batch_normalization_2 (Batch	(None, 128)	512
dense_3 (Dense)	(None, 64)	8256
batch_normalization_3 (Batch	(None, 64)	256
dense_4 (Dense)	(None, 32)	2080
batch_normalization_4 (Batch	(None, 32)	128
dense_5 (Dense)	(None, 16)	528
batch_normalization_5 (Batch	(None, 16)	64
dense_6 (Dense)	(None, 1)	17

Total params: 210,369
Trainable params: 208,353
Non-trainable params: 2,016

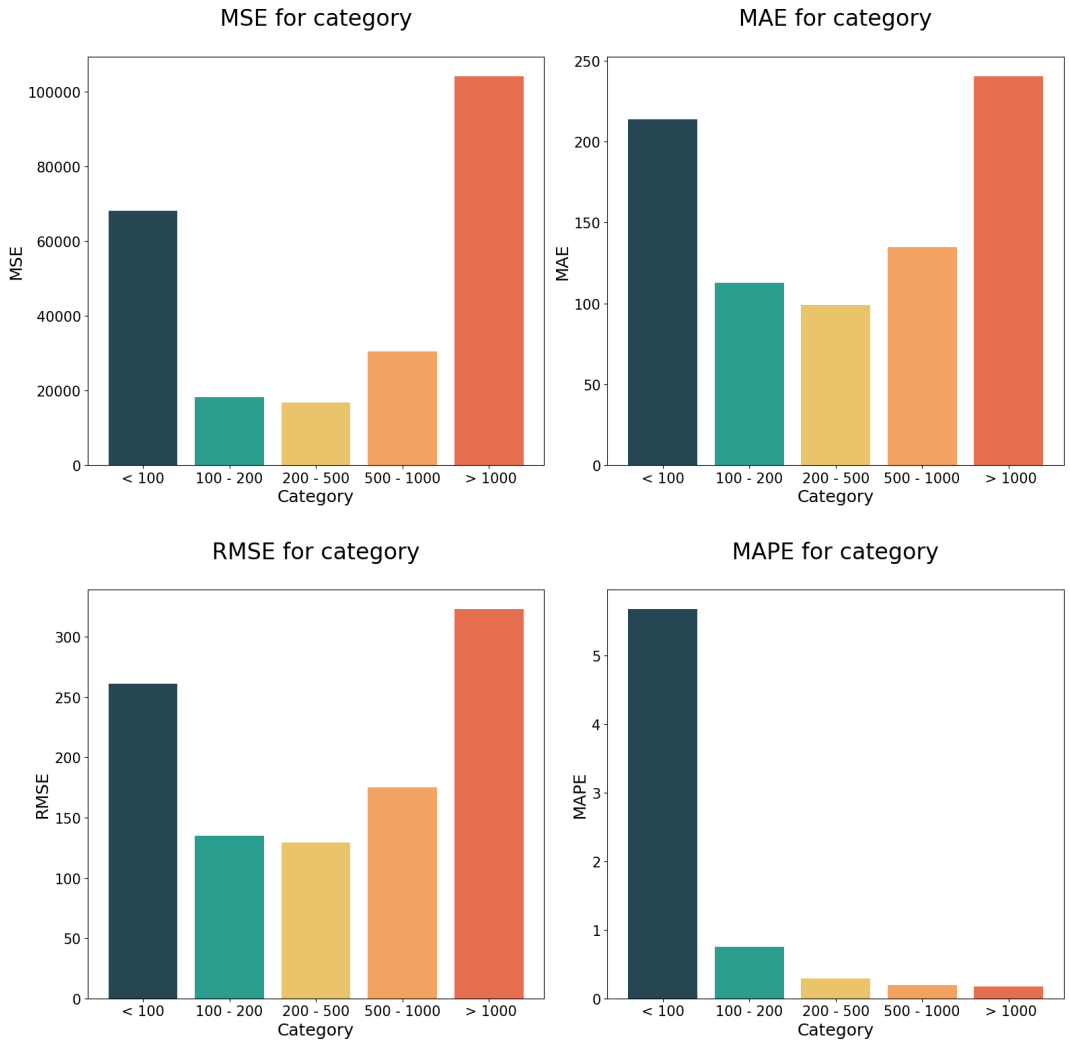
Results

The following graphs represent the loss function of the model made on the training set respect to the one made on the validation set.



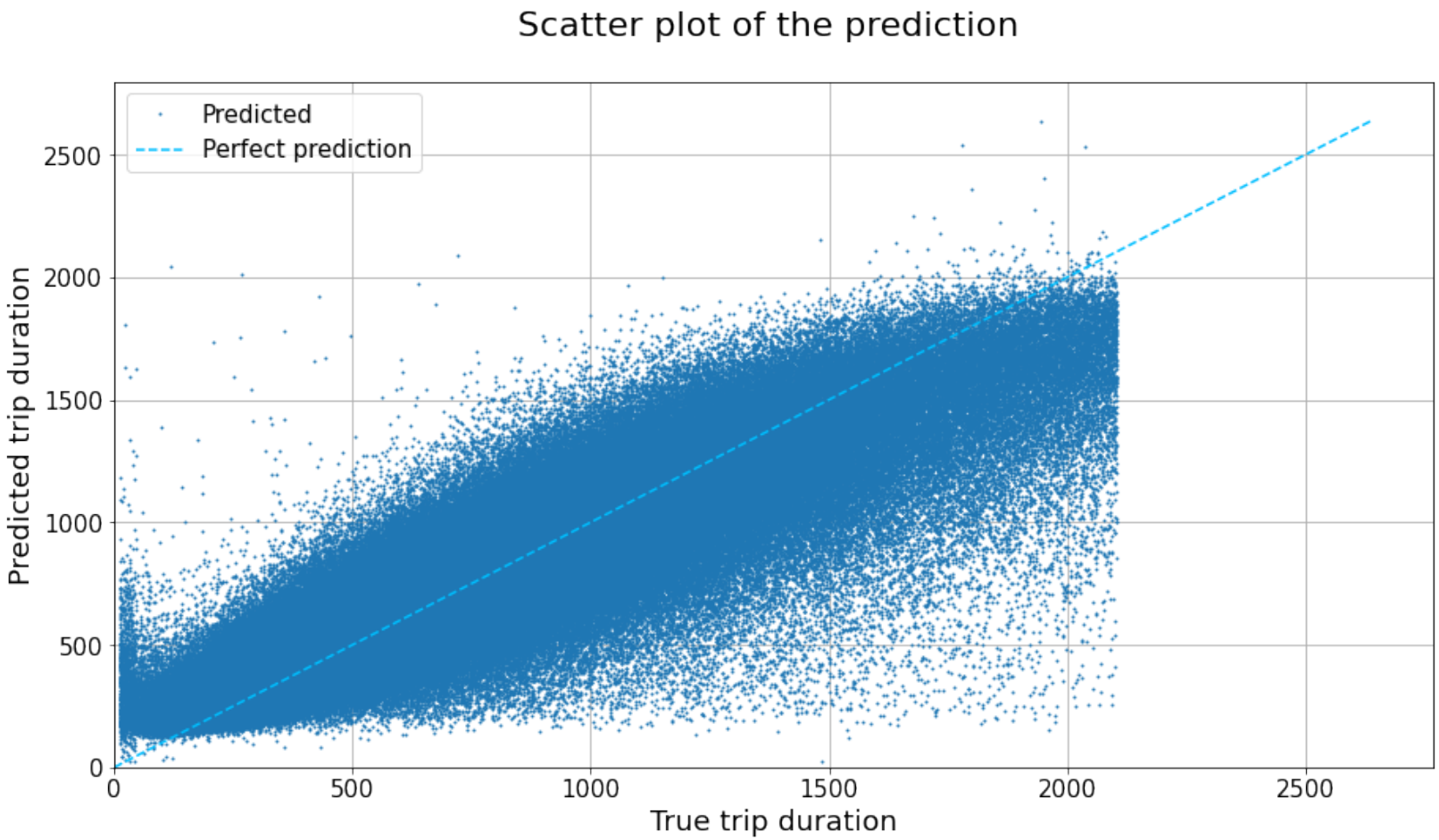
Results

The following graph shows the error divided by trip duration's bin category



Results

The following graphs represent the scatter plot of the prediction against the real duration.



Conclusions

The approach applied gives good results.

In particular, the integration with geo-spatial data is important in order to consider different types of trips and also the outlier cut of too long trips and too short to be meaningful.

The models tried are quite complex but the best is the simplest one (5 layers instead of 7 layers),usually a simpler model avoids overfitting and is more robust.

The project could be improved in the future by adding some traffic information or based on this model bring a real-time one, that can be much more useful.

Image from https://fortune.com/2016/07/13/study-nyc-taxi-outperforming-uber/

kaggle

Thank You

Christian Uccheddu - 800428
Federico Luzzi - 816753
Federico De Servi - 812166

Image from https://fortune.com/2016/07/13/study-nyc-taxi-outperforming-uber/