New York Taxi project

Goal: we want to predict the avarage amount of money spent on a casual taxi ride in NYC for each region per given day & hour so taxi drivers will know when they should go.

#1 step - Data exploring:

Understand what features we can "let go", finding "odd" information (like negtive payment)

See is all features are in the best data type

start thinking on features we want to add.

We had 19 features - some of them were unrelevent

We found some threshold for expensive rides to delete.

2 step - Cleaning & preperation

Clean & preper base on exploring.

we cleaned all "\r"

change datatypes of dates, and locations IDs to string

cleaned - Not Jan 2024 and pyments

Prepring: collecting all rides at the same time and loction by the mean. of total amount
adding # of rides

#3 step - Benchmark model

Test our features on a simple model to find problems. and have some comparison point.

Chose Decision tree regression

tried two depths - both seems to have

a problem predicting small numbers.

Features used

Pickup location
trip date - day + hour
total amount (mean)
rides

Problems

We probably need to add more features

Ideas - weekend - yes/no
holydays
rush hours
late night connects to early mornings

maybe search online for more ideas.

need to find out where the small vals are coming from. maybe observe rides under 5$

the model never predict high values

#4 step - feature engineering:

fix + add features using the conclusions on step 3.

added - weekday (0-6)
is weekend
sin + cos of hour to connect late night & early day

delete - any rides under 1 dollar.

#5 step - Model training

Try new models & benchmark after "fixing" the features

To do