Last Judgement Report

A.K.J.G. de Wit

Maastricht University

1 Problem Statement

For this exam it is asked to predict the place of impact for a meteorite. A dataset that is not complete untill the meteor hits the ground is given, as well as a dataset of buildings and their residents in Maastricht. Using models that were taught in the Data Analysis course it is attempted to estimate the point of impact, and therefore estimate how many people will be affected by the meteor.

2 Introduction and Description of Data

The problem of predicting where a meteorite might land requires the information of its position during a certain time period. This information was given in *'radar.csv'*, containing x, y, and z-coordinates in kilometers as well as time in seconds. Because the file only contains 130 samples it was no challenge to comb trough the file to look for missing values, plot the values and look for extraordinary values. After careful inspection and plotting of the initial data it could be concluded that the data was clean and did not need any treatment.

The file containing data about the buildings in Maastricht, *'mbuildings.csv'*, did have some missing values for the 'residents' column. This was solved using the build in pandas function *'interpolate'* using a spline to fill in the missing data according to the amount of bed-, and bathrooms.

With both files checked and filled where necessary, the next step of plotting the data could be started.

2.1 Plotting the data

A logical step when approaching the stated problem is to plot the data to see what is happening. The result of this can be seen in Fig. 1. It is immedeately obvious that the meteor is falling as seen in the 'z' axis. While it is also moving over the horizontal plane according to the 'x', and 'y' axis. With this plot one can get a good intuition of where the meteor might strike using human reasoning. But to get more precise, models will be needed.

3 Modeling Approach

Judging by the graph that was plotted in Fig 1, a simple linear regression model might suffice to predict what the trajectory of the line might be. Using the *LinearRegression* package from sklearn without any parameters the baseline model was created. These results can be seen in Fig 2. The linear regression seems to be a good basic fit, however looking at the root mean square error and the r2 score in Table 1 it is evident that this model is too simple.

Because this is a regression problem and the linear regression already did quite well for such a basic model, the next choice was to explore polynomial linear regression. Using the sklearn function *'PolynomialFeatures'* to preprocess our time data into more dimensions. "For example, if an input sample is two dimensional and of the form [a, b], the degree-2 polynomial features are $[1, a, b, a^2, ab, b^2]$"[1]

After the time has been preprocessed for our polynomial linear regression the model is fit again and we see quite the improvement compared to linear regression. The plots can be found in Fig 3 and their respective error scores in Table 1. This model was considered robust enough for the problem at hand, so the next step was to tune the model

4 Results and Interpretation

Q1 Estimate point of impact

"Using (any) methods you learned in this wonderful class to estimate the expected point of impact."
After deciding on polynomial linear regression for the model a couple of parameters could be tuned to find the best possible model. Using a for loop to change the degrees of the polynomial preprocessing and inside that a GridSearch to find the best parameters for the linear regression. The result was that a degree of 2 was the best with with two parameters to be tuned in the linear regression. (*See Table 2*)

After the model was tuned to be as close to perfection as it could a new list of 'future' times was made and the predictor was applied to this future time series. The plots of extending the meteor trajectory can be seen in Fig 4. With this prediction of the trajectory based on the earlier list of different times an estimation of the meteor's location could be made up to 750m off the ground.

To get as close as zero as possbible an adaptation of the bisection method was used to estimate the exact time of the meteor touching the ground. With this information, the trained models for each coordinate point could be passed this time and the result is the (predicted) point of impact:

$$x = 7047.744Km \; y = 6589.909Km \tag{1}$$

Q2 Area of impact

. **"Provide the Gemeente with not just an impact point but an area that has to be evacuated"**
To find the area for the impact instead of just one point, the resampling method

[1] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html

that is described in the exercise was used. By using a for loop that replicated the exact process of Q1 with the resampled datasets, a list of 10.000 points of possible impact was created. *(See Fig 5).*

Using Gaussian Kernel Density Estimate[2] to get the density of the predicted points it was easy to get the points that had 90% of the maximum density.

This area (Fig 6) is the area that is 90% likely to get hit by the meteor.

Q3 People in danger

"Using the building database, estimate the total number of people that will most likely be affected. Use the region you defined in Q2."
To find the buildings that are situated under the area found in Q2 a function was made to create a range around every point that was inside the impact zone. With this range given, the only thing to do was check the buildings that had a 'x' and 'y' value inside one of the impact points ranges, and if so conclude that this building would be hit. *(See Fig. [7, 8])*

Then using the interpolated data of residents and the buildings that potentially would be hit, an estimate of people that will most likely be affected can be made. And the model gave the estimate of:

$$79\ people \tag{2}$$

Q4 More accurate readings

"Provide an answer for the impacted region and comment on the result."
All the steps that were taken to get to the answer of Q3 were used again given the new dataset. This resulted in a completely new possible area of impact for the meteor and increased the number of possible affected people to 262. To see the difference between the two impact areas, see Fig [9, 10]

5 Conclusion

The results achieved in this exam all come from a linear regression model, given that the data is preprocessed to make the model a polynomial linear regression, it is still a very simple way of regression. This simplicity does make it robust to changes and new data as we have seen in Q4, however it feels like there should be better models to predict the trajectory of a falling object. Another interesting thing would be combining the two sets of data from the two different radars to see if that will give even more precision to the predictor.

[2] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.html

Appendix

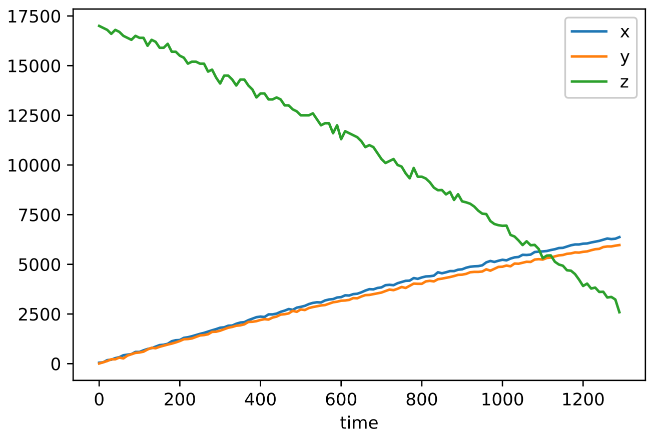
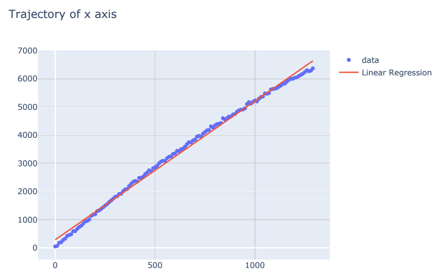


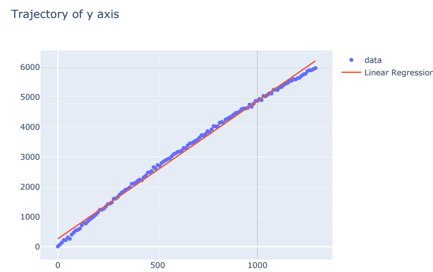
Fig. 1: The coordinates of the radar.csv file plotted over time.

Table 1: RSME and R2 scores for different models.

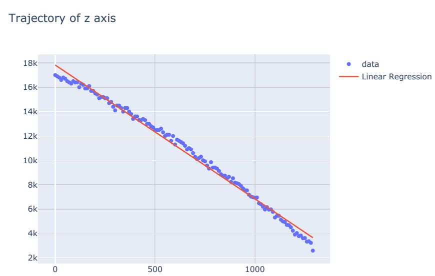
Model	RMSE sum	R2 sum
Baseline LR	624,375	2.982
Polynomial LR	214,818	3



(a) x-axis



(b) y-axis

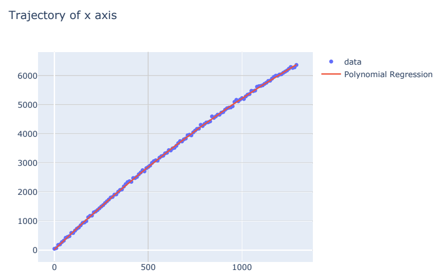


(c) z-axis

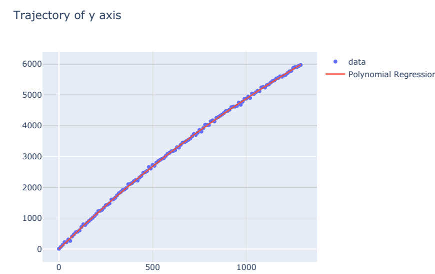
Fig. 2: Linear Regression baseline

Table 2: Degree and gridsearch results.

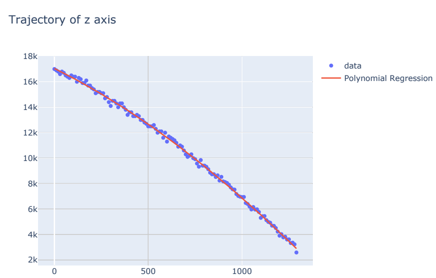
Degree	RMSE sum	R2 sum	Parameters
0	13578.080	-1130.513	'fit_intercept': True, 'n$'_{jobs} : None,' normalize' : True$
1	1458.214	-8.85	'fit_intercept': False, 'n$'_{jobs} : None,' normalize' : True$
2	213.652	2.809	'fit_intercept': True, 'n$'_{jobs} : None,' normalize' : False$
3	202.840	2.845	'fit_intercept': False, 'n$'_{jobs} : None,' normalize' : True$
4	267.236	2.627	'fit_intercept': False, 'n$'_{jobs} : None,' normalize' : True$
5	310.308	2.699	'fit_intercept': True, 'n$'_{jobs} : None,' normalize' : True$



(a) x-axis

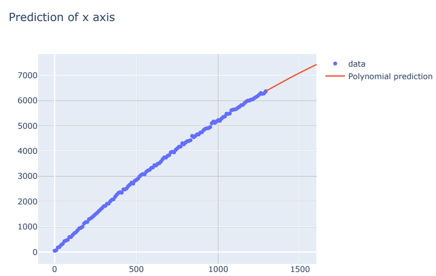


(b) y-axis

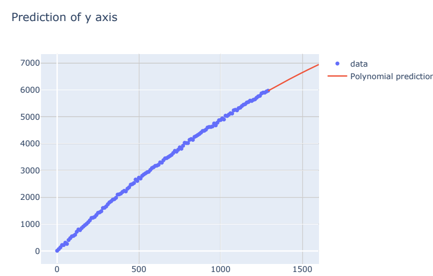


(c) z-axis

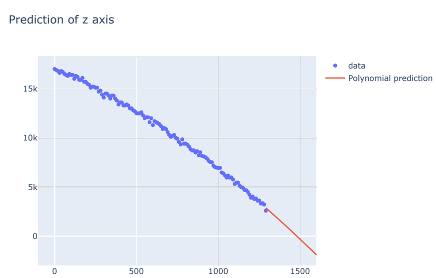
Fig. 3: Polynomial Linear Regression



(a) x-axis



(b) y-axis



(c) z-axis

Fig. 4: Prediction

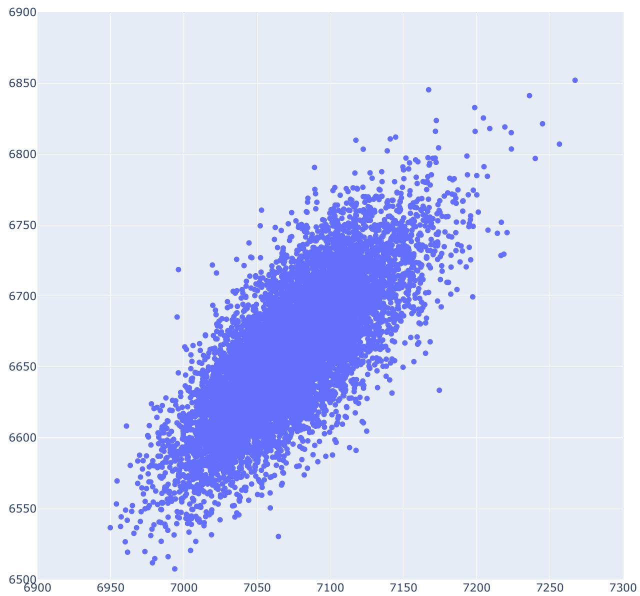


Fig. 5: 10000 points of possible impact.

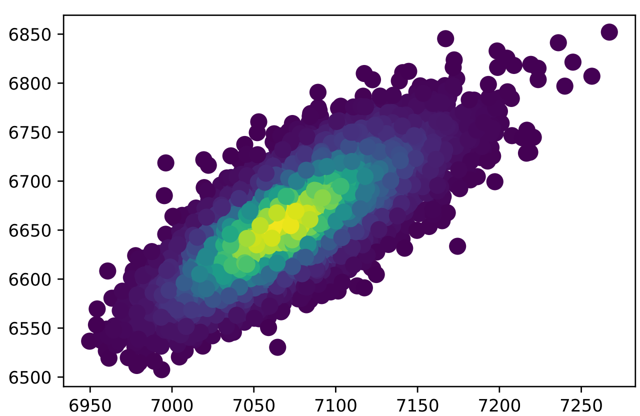


Fig. 6: Point density.

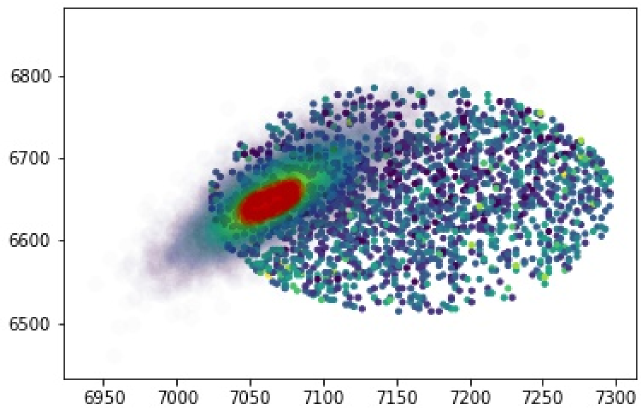


Fig. 7: Buildings and impact area.

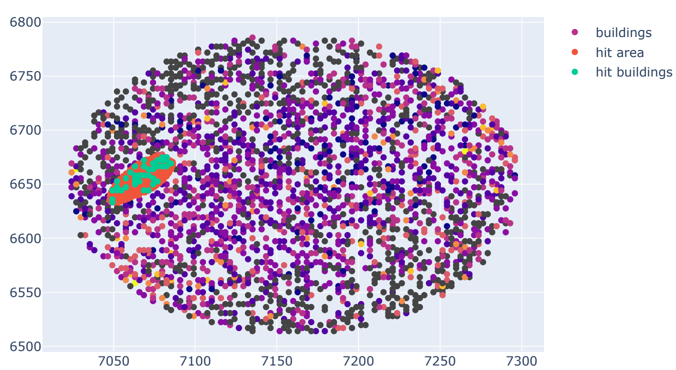


Fig. 8: Buildings and impact area.

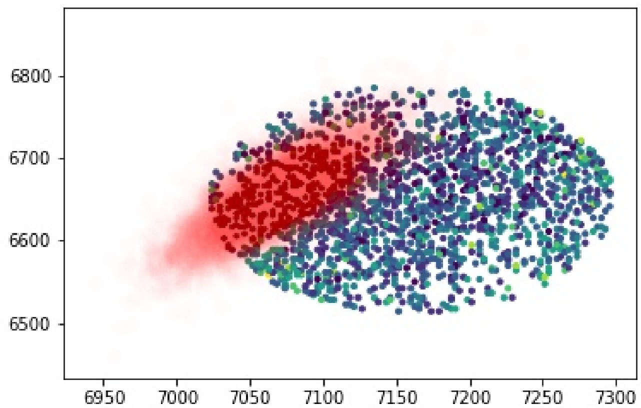


Fig. 9: Radar impact on buildings.

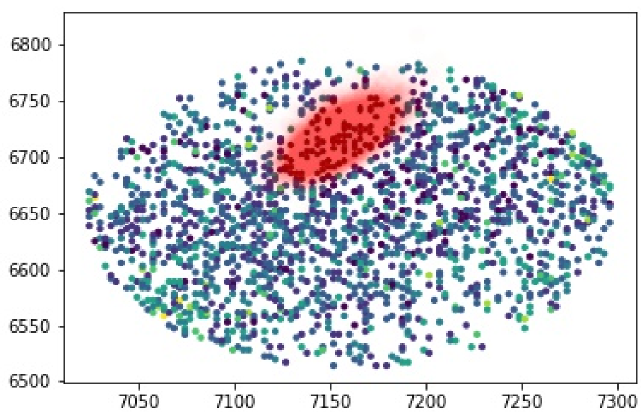


Fig. 10: Radar2 impact on buildings.