The Deluge – Flood Risk in the UK

This project had two aims:

1. To create a risk tool to predict local authority, median house price, flood probability and overall flood risk.
2. To create a visualisation tool to represent this information.

Median Price Model

KNNeighbors was chosen, with n_neighbors=5. Different models were compared as seen in Table 1 (scores in log form), with KNN performing the best. The best KNN parameters were selected through grid search. The models were trained on the log form of the price as the median price distribution was very unbalanced.

The sample density distribution was very uneven, with many more samples in London than elsewhere. To solve this problem, we tried to cluster the data using K-means and run separate models on each cluster. This did not improve the performance. We also tried to divide our data based on soil type before applying KNN, as the high density of samples correlated with soil type of Unsurveyed/Urban. This did not improve the performance.

The important features for the model were Easting and Northing. Adding other features worsened the prediction.

Local Authority Model

KNNeighbors was chosen because the input values can represent actual distances. A grid search was used to tune the n_neighbors parameter and n_neighbors = 1 gave the best score. The score was ~97% so no other models were needed. The only features used were Northing and Easting.

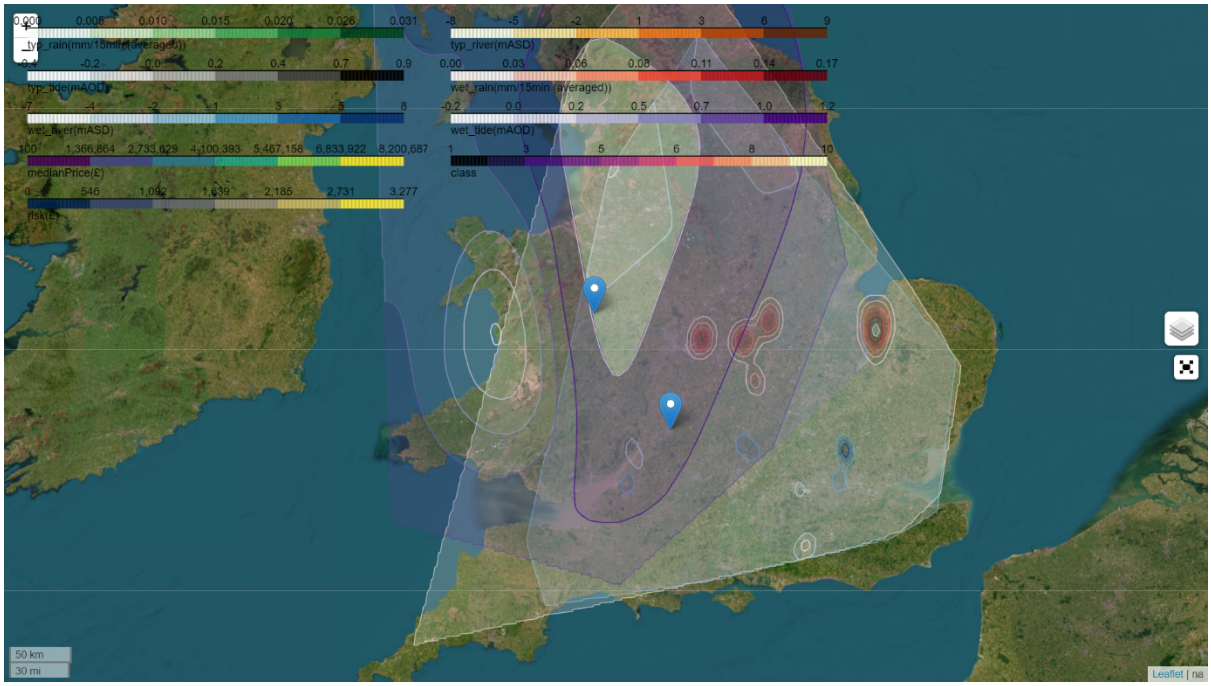
Flood Probability Model

Random Forest regressor and KNN regressor were chosen after testing all other algorithms, as seen in Table 2. For feature selection, easting, northing, altitude and soil type were selected because the other features were deemed as non-influential in predicting the risk label/flood class.

Annual Flood Risk

Overall flood risk is calculated according to the formula R = 0.05 x (total property value) x (flood probability). To make this prediction we calculate total property value using the median price model. We then combine this with the result from the flood probability model.

Visualisation Tool



The red region is the average rainfall data, blue region is the average river data, and the purple region is the average tidal data on a wet day. The stations with stationReference 2830 and 2660 are predicted as risky areas because the river_level is higher than the typicalRangeHigh.

- For station 2830: river_level: 0.984, typicalRangeHigh: 0.97.
- For station 2660: river_level: 0.063, typicalRangeHigh: 0.05.

APPENDIX

Table 1: Model comparison for the median house price estimate.

Model	Features	RMSE	R^2
Dummy model strategy="mean"	easting, northing, localAuthority, altitude, soilType	0.7638	-9.7707e-06
Linear Regression	easting, northing, localAuthority, altitude, soilType	0.6076	0.3671
SGDRegressor/SVR/DecisionTreeRegressor/ AdaBoostRegressor/GradientBoostingRegressor/ XGBRegressor	easting, northing, localAuthority, altitude, soilType	0.60~0.75	0.30~0.48
KNN {'algorithm':' ball_tree', 'leaf_size':30, 'n_neighbors':15, 'p':2, 'weights': 'distance'}	easting, northing	0.5126	0.5496
Kmeans + KNN K=15, n_neighbors=5	easting, northing	0.5225	0.5320
KNN (split based on soil type) n_neighbors=5	easting, northing, type('Unsurveyed/Urban': 1, else:0)	0.6948	\
Bagging (KNN) n_neighbors=5	easting, northing, type('Unsurveyed/Urban': 1, else:0)	0.6878	\
Stacking (LR, RF, KNN)	easting, northing, type('Unsurveyed/Urban': 1, else:0)	0.7485	\

Table 2: Model comparison for the flood probability estimate.

Model	RMSE	MAE
RandomForestRegressor()	1.172	0.427
XGBRegressor()	1.203	0.538
BaggingRegressor()	1.213	0.428
KNeighborsRegressor(n_neighbors=10)	1.218	0.481
ExtraTreesRegressor()	1.222	0.412
GradientBoostingRegressor()	1.285	0.627
MLPRegressor()	1.297	0.627
LinearRegression()	1.375	0.759
SGDRegressor()	1.378	0.748
AdaBoostRegressor()	1.543	1.196