From Prototyping to Production:

Predicting Mobility Patterns using Machine Learning

Marvin Hedderich

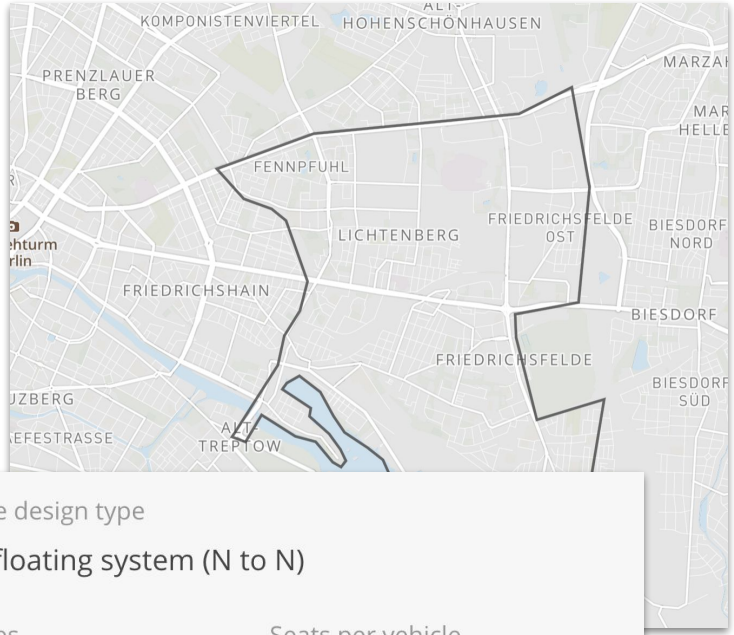
door2door

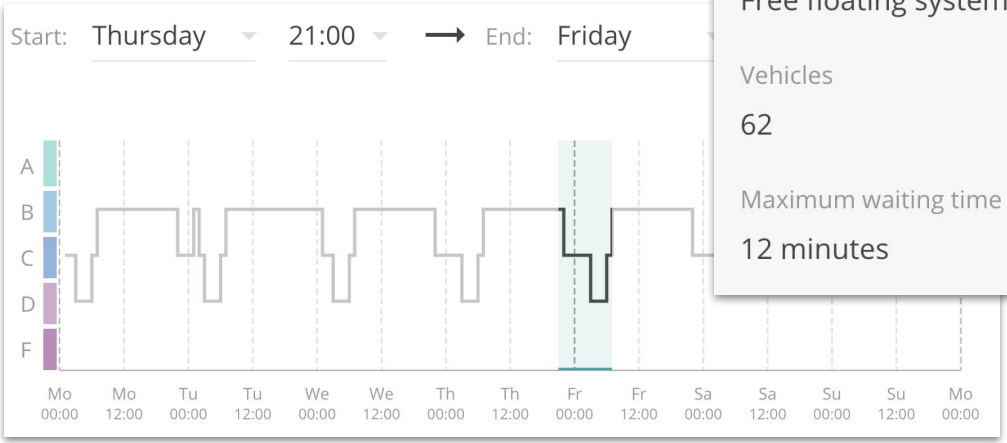
Planning & Operating a Ridepooling Service

1

How to plan a Ridepooling Service

- Operating area
- Operating period
- Service parameters





Service design type

Free floating system (N to N)

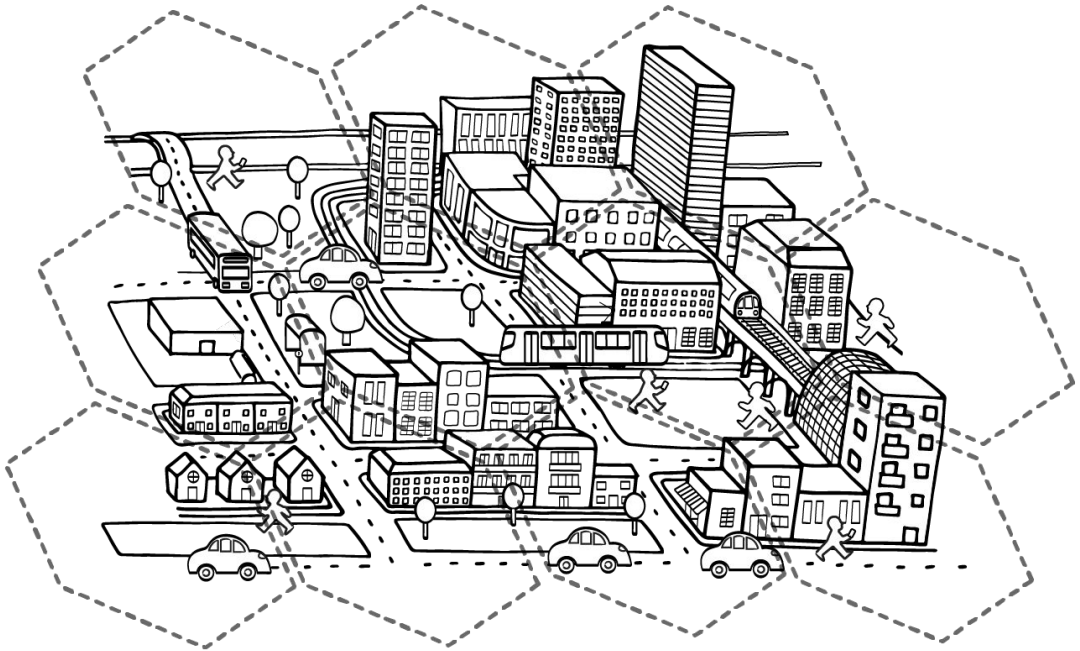
Vehicles	Seats per vehicle
62	5
Maximum waiting time	**Maximum detour time**
12 minutes	10 minutes

The Prototype

2

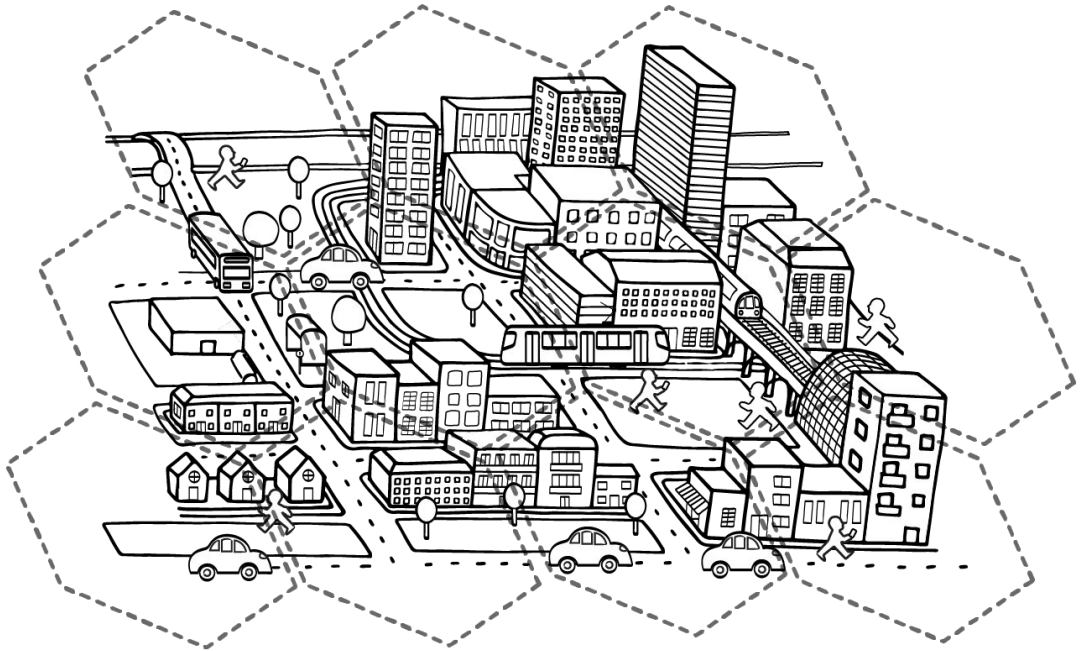
Dividing the City into Chunks

- Hexagons of equal size
- 250 metres radius
- Prerequisite for scoring and connections



Defining Features

- Population density
- Residential buildings
- Shops, offices
- Restaurants, bars, nightclubs
- Cinemas, theaters



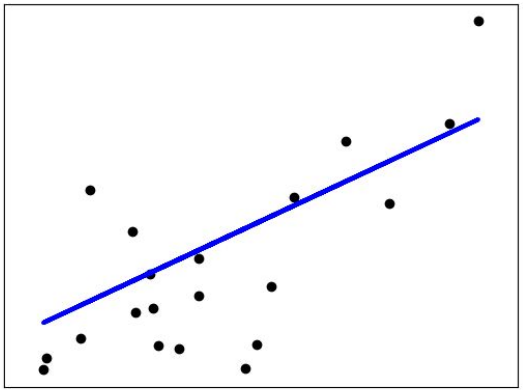
Training Dataset Example (Spatial Distribution)

cell_id	population_density	office_buildings	sustenance	demand
cell-1	5483	2	9	174
cell-2	8420	14	11	766
cell-3	2370	4	5	219
cell-4	11743	7	18	1764

Training and Prediction

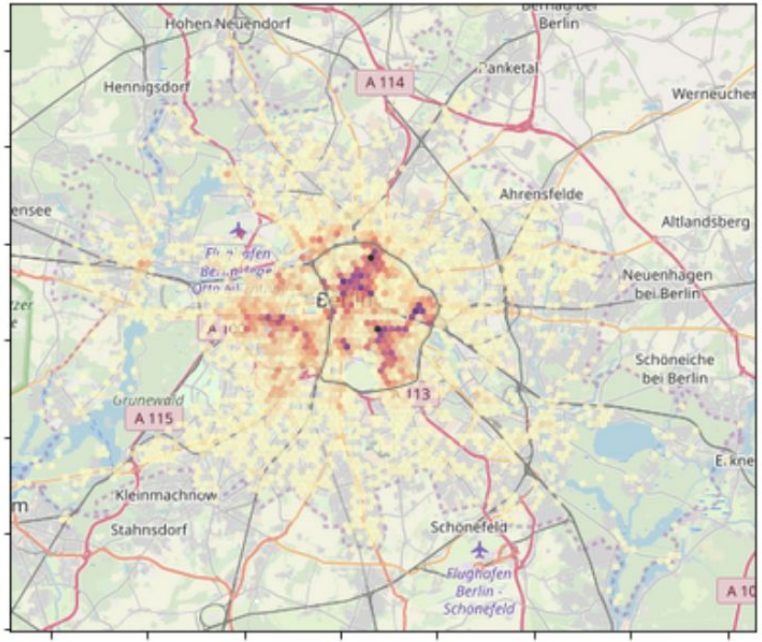
```
from sklearn import linear_model

model = linear_model.LinearRegression()
model.fit(training_region_X, training_region_y)
predicted_demand = model.predict(prediction_region_X)
```

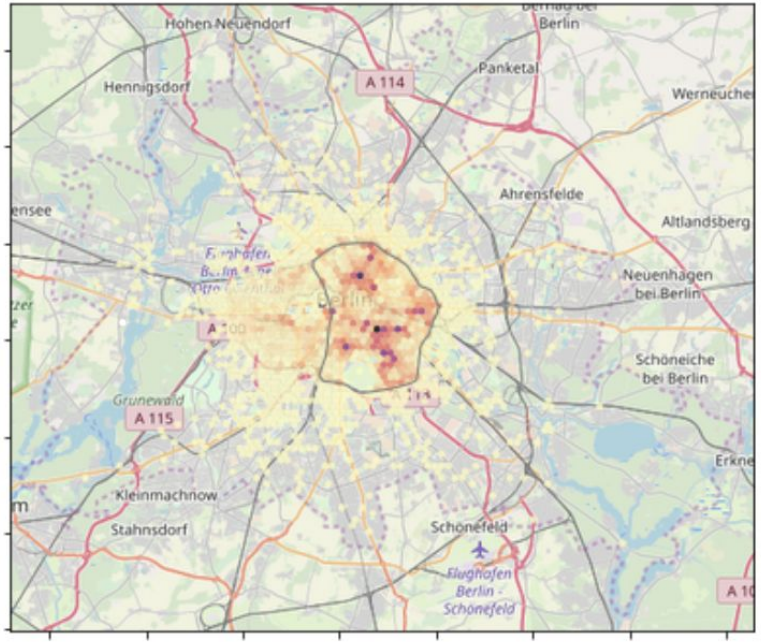


Results 🎉

Predicted Rideshare demand

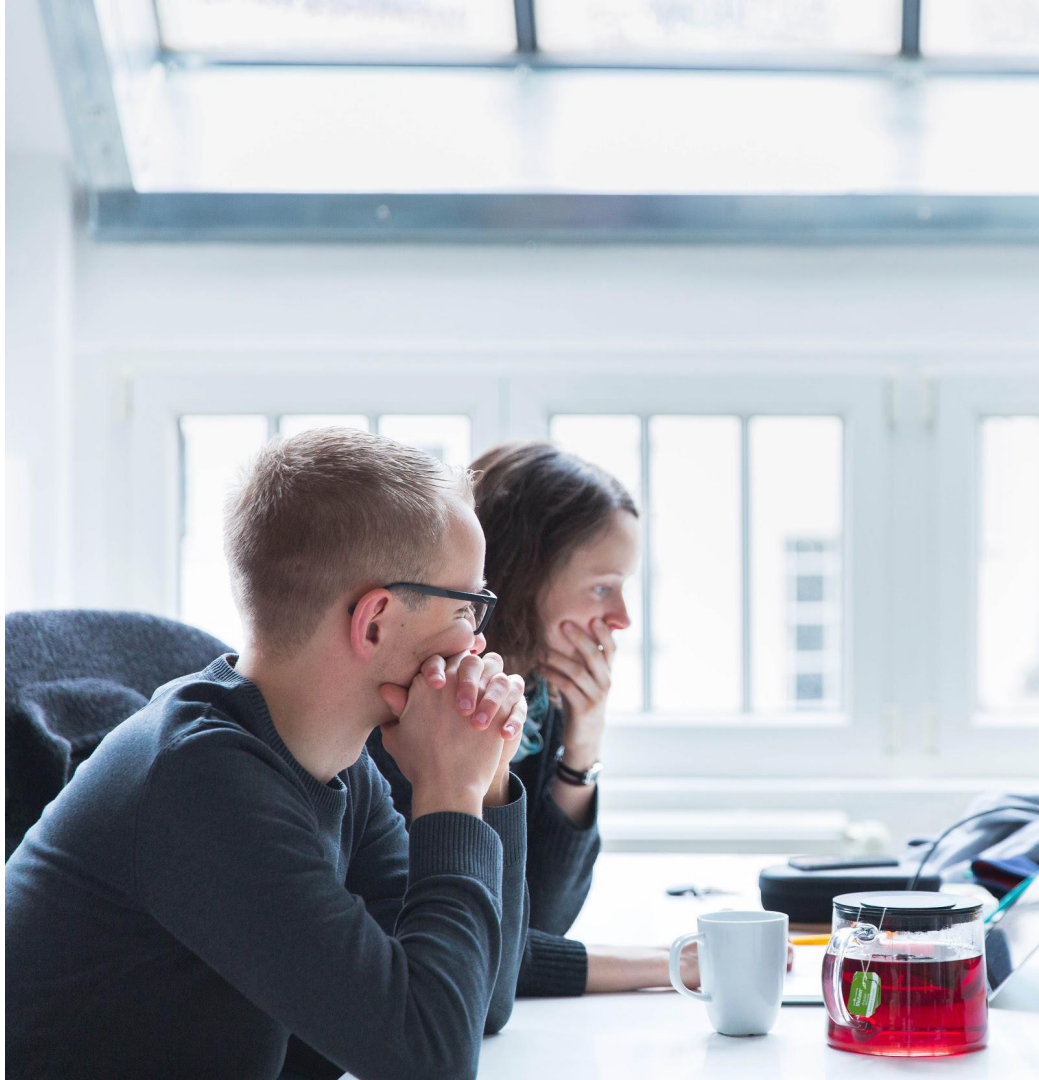


Actual Rideshare demand



Learnings

- Predicting spatial distribution works!
- Comparing algorithms is easy
- Visualization is a must
- Feature importance



Limitations

- No connections (“trips”), only spatial distribution
- No temporal distribution
- Accuracy?

Introducing Connections

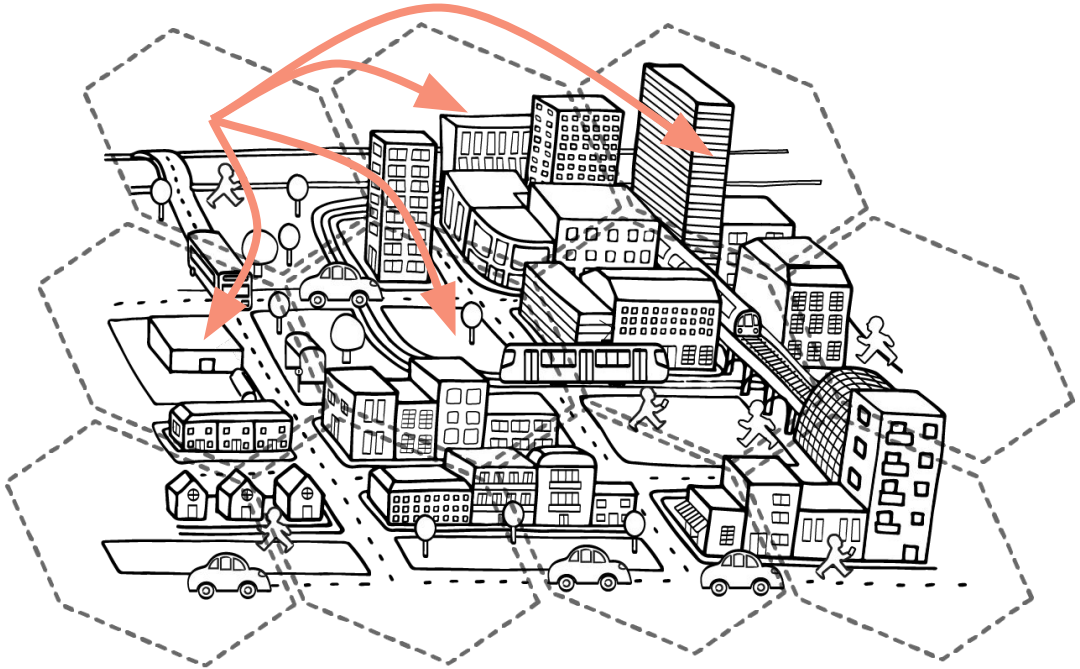
3

Introducing Connections

- Every cell possibly has trips to every other cell.

  11 cells → 121 connections

- Origin-Destination Matrix



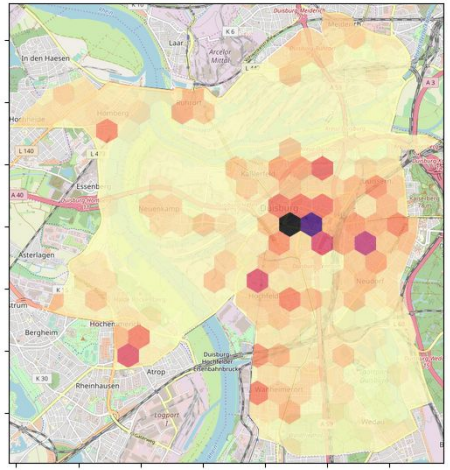
Training Dataset Example (OD-Matrix)

cell_from	cell_to	office_bldgs_from	office_bldgs_to	...	weekly_trips
cell-1	cell-1	2	2		2.2
cell-1	cell-2	2	14		72.6
cell-1	cell-3	2	4		12.0
cell-1	cell-4	2	7		23.9

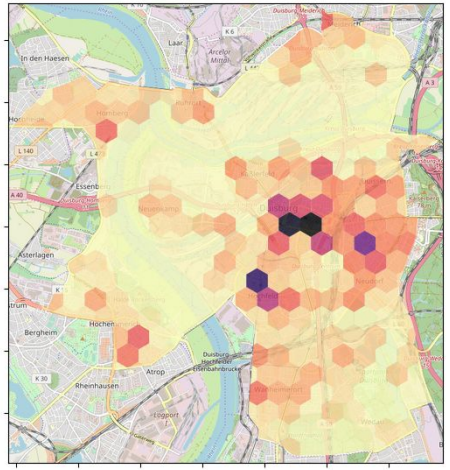
Predicting Connections

- Predicting OD-Matrix based on our set of features seems possible
- Estimation of OD-Matrices based on population data was researched before

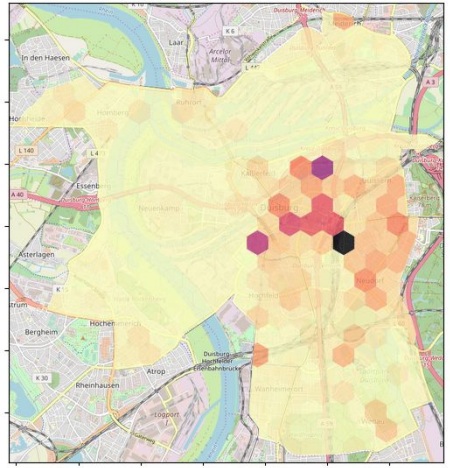
Predicted Rideshare demand (to)



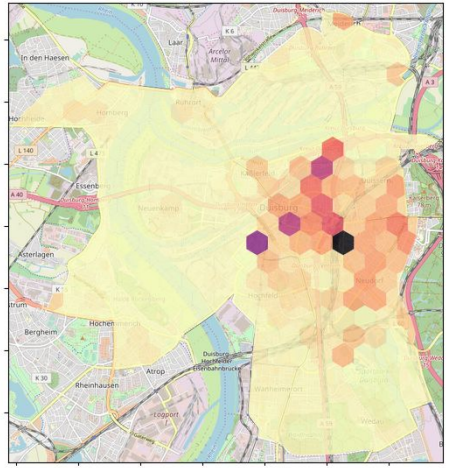
Predicted Rideshare demand (from)



Actual Rideshare demand (to)



Actual Rideshare demand (from)



Scaling the Data

4

Scaling the Data

- Different regions have different feature densities
- Scale values relative to other cells of the same region

Measuring Success

5

Defining Accuracy

- Measure based on RMSE
- Normalization to scale from 0 to 1
- Accuracy:
  1 - normalized RMSE

Root Mean Square Error

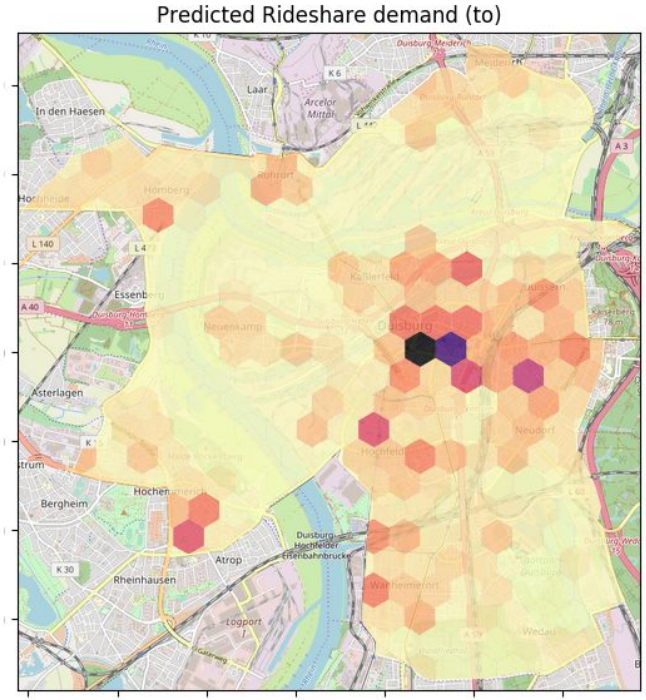
RMSE used to aggregate deviations across all values between predicted and actual datasets, weighting outliers

Testing for Accuracy

Method	Accuracy
Linear Regression	98%
Polynomial Regression	98%
Random Forest Regression	98%
Neural Network	98%

Re-Defining Accuracy

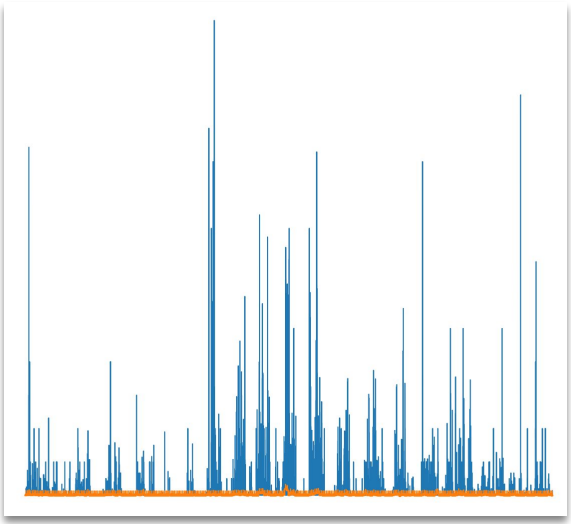
- Measure based on RMSE
- Normalization to scale from 0 to 1
- Accuracy:
  1 - normalized RMSE
- Exclude zero values



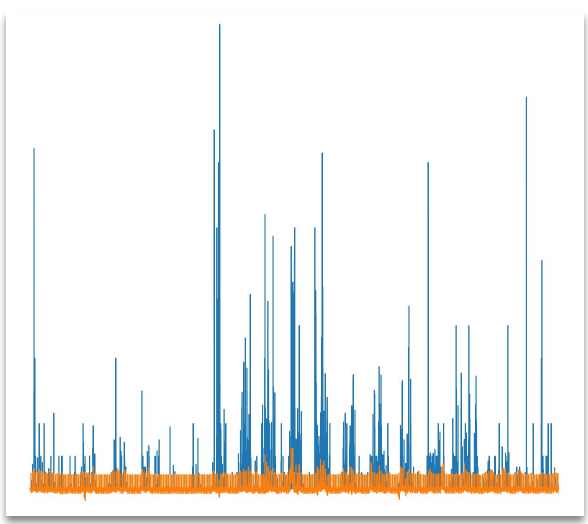
Excluding Low Values

	Error (absolute)	Accuracy (%)	Connections compared
Full	0.163	98.82	69696
Top 50%	0.226	98.38	34848
Top 10%	0.477	96.59	6970
Top 1%	1.469	89.50	697
> 0 (in actual)	0.787	94.37	2515
> 0.5 (in actual)	1.478	89.43	688
> 1 (in actual)	2.269	83.79	218
> 2 (in actual)	3.789	72.92	57
> 5 (in actual)	8.388	40.08	5

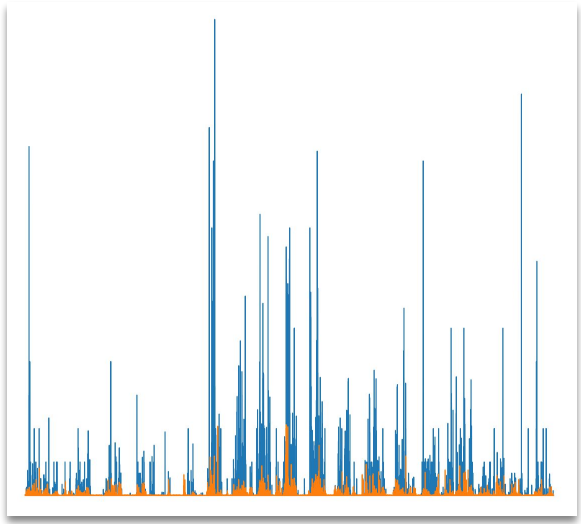
Visualizing Data: Value Distribution across Connections



Linear Regression

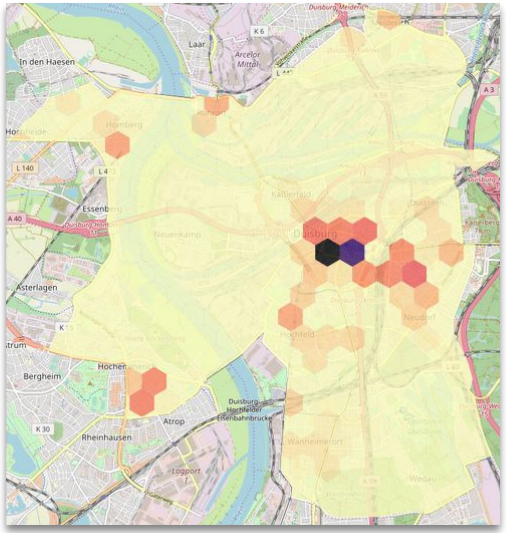


Polynomial Regression

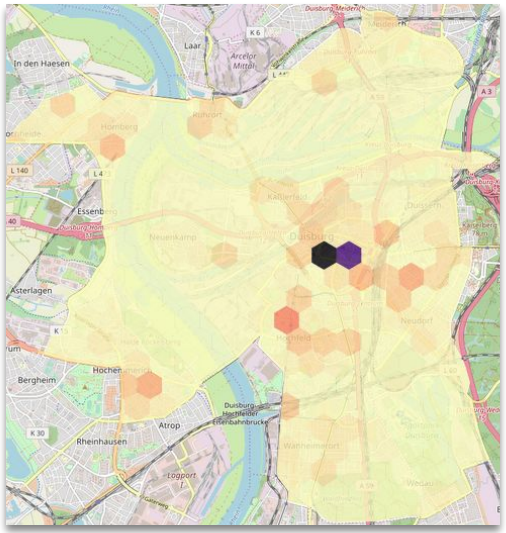


Random Forest Regression

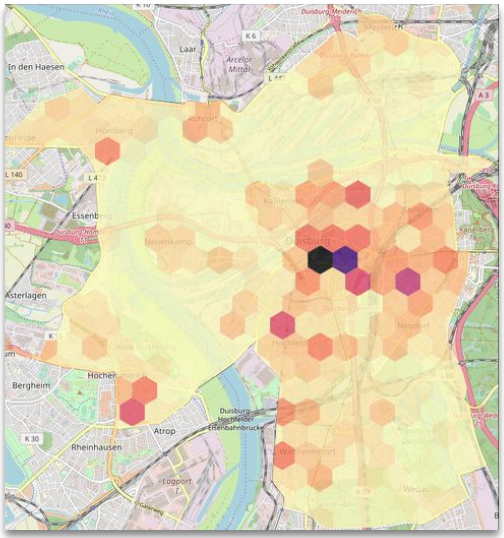
Visualizing Data: Spatial Distribution (Destinations)



Linear Regression



Polynomial Regression



Random Forest Regression

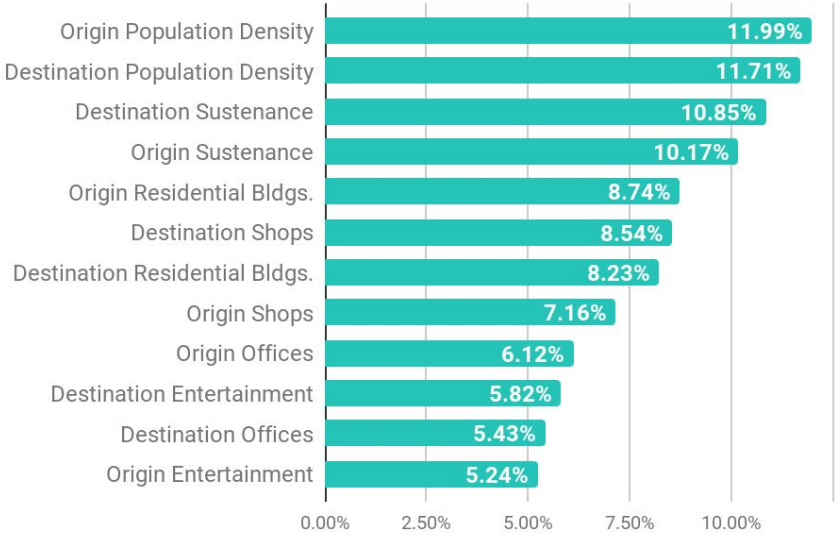
Finding the Right Features

6

Finding the Right Features

```
from sklearn.ensemble import RandomForestRegressor

model = RandomForestRegressor()
model.fit(training_region_X, training_region_y)
model.feature_importances_
```



Learnings

- Know your numbers
- Visualize as soon as possible
- Predicted patterns are only as good as your data

Bremen

Region: Bremen
Area: 317.85 km² Inhabitants: 551,632
Time: From Friday at 17:00 to Saturday at 01:00

More Details

Analysis

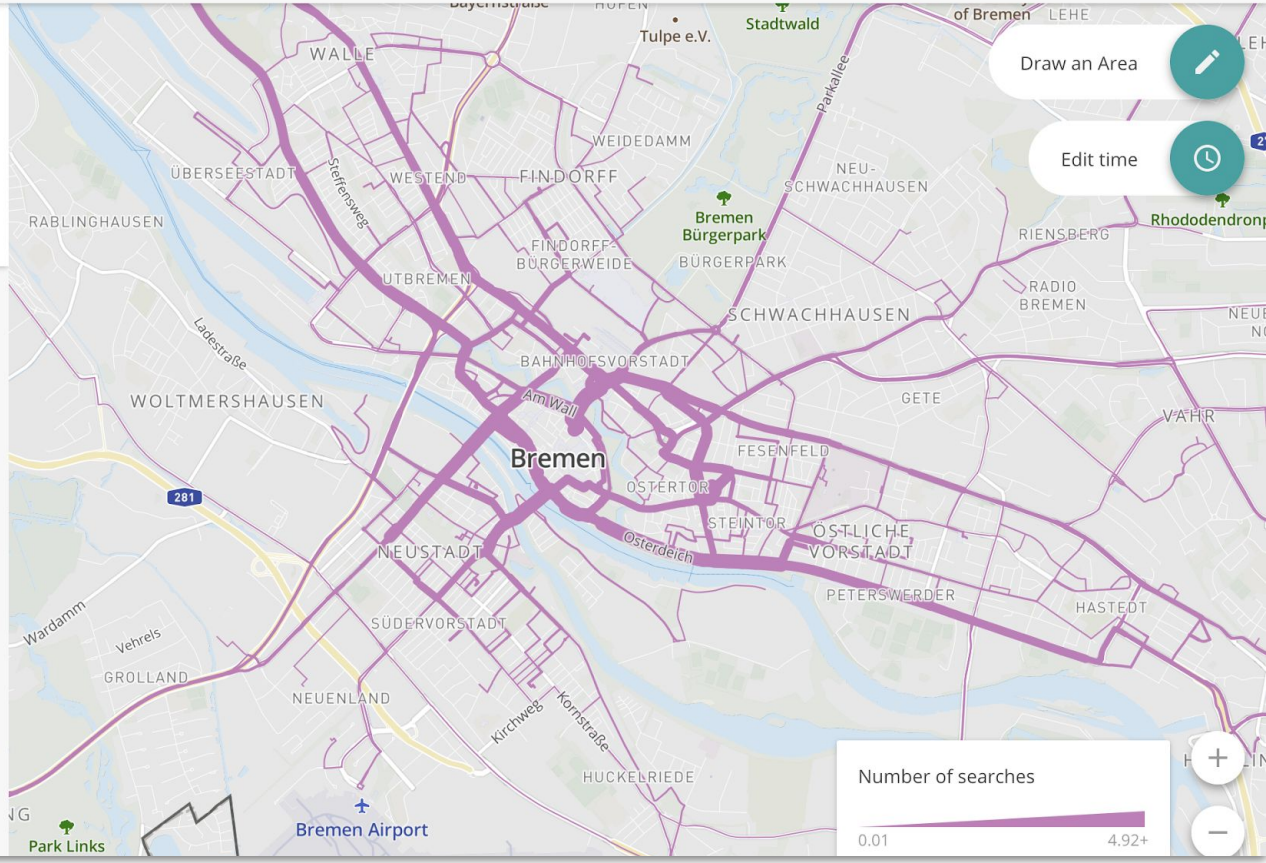
Supply Demand C

Pooling Potential
60%

Required
235
bookings/h

Supply Indicators

Walking Accessibility
ø 1.26 stations Medium



Join the ride!

people@door2door.io

THANK YOU!

@door2doorHQ

@door2door_team

door2door.io