Crowdedness Prediction Model

Don de Lange
Msc Data Science

Supervisor: Noud de Kroon

Introduction

- Amsterdam is crowded
    - High pressure public transport
    - More pedestrians
    - More guests public events
- Municipality of Amsterdam → Predict future trends

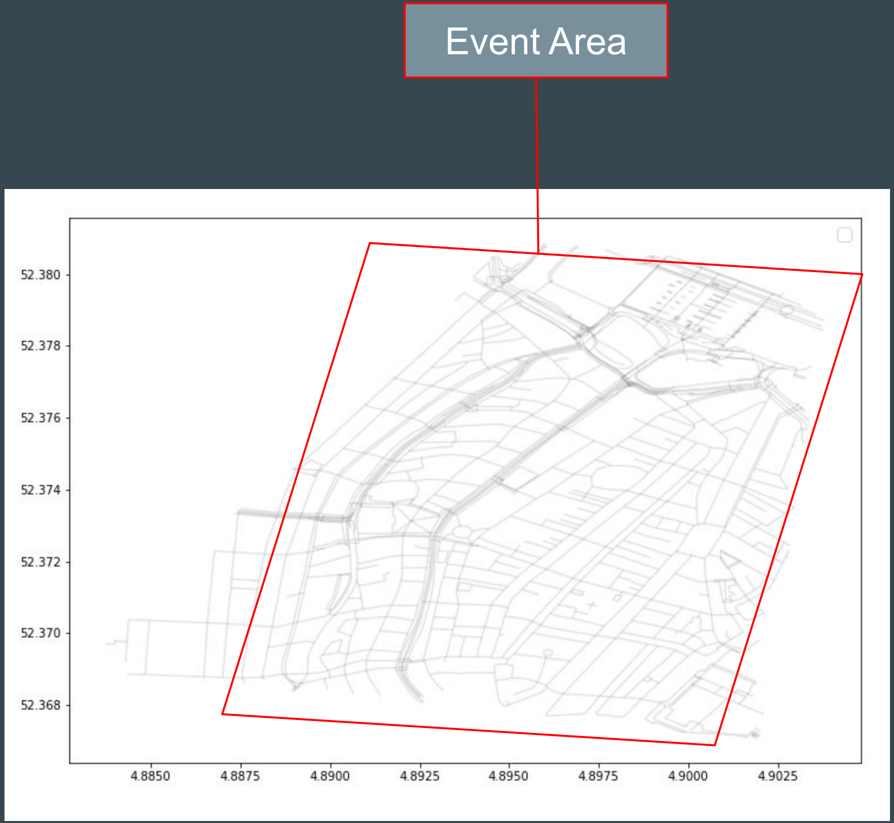
How can a prediction of crowdedness within the city of Amsterdam be given, based on input from city-wide available data sources?

Data

- Start Date: *11 March 2018*
- End Date: *24 March 2019*
- Measurements made per day, per hour

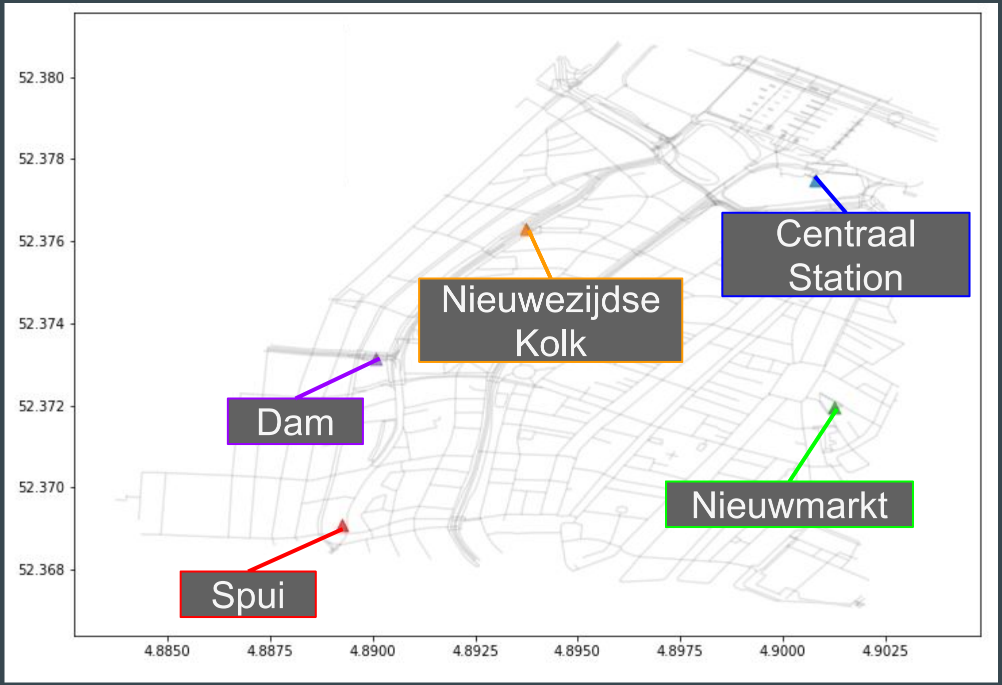
Event Dates

- Categorize dates with events as outliers



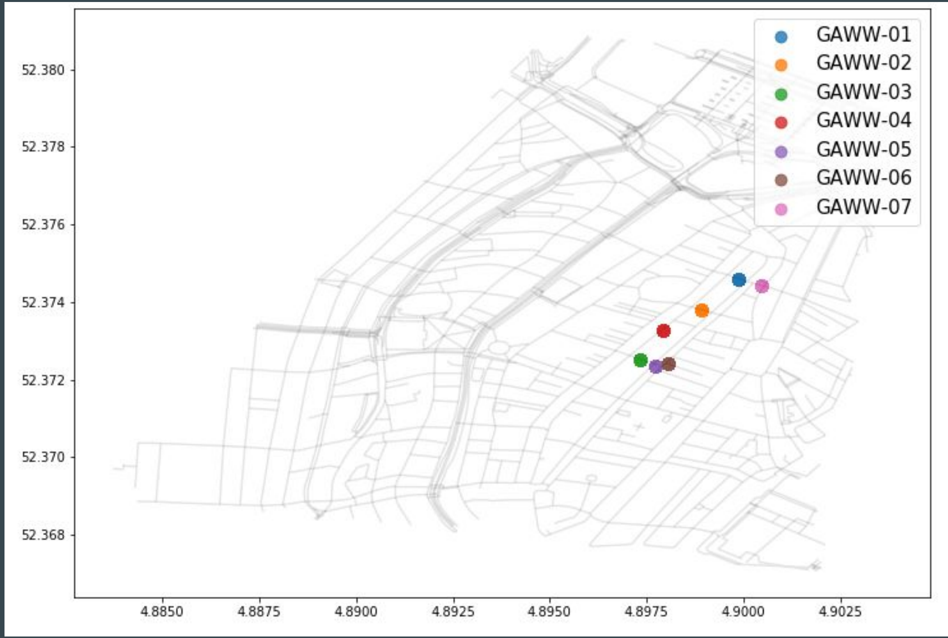
Public Transport

- GVB
- Per Station
  - Number of Passengers
  - Co-ordinates



Crowdedness Sensors

- CMSA
- Sensor
  - Street Zone
  - Sum counts made with Count Cameras and Wi-Fi sensors
  - Co-ordinates
- Missing values



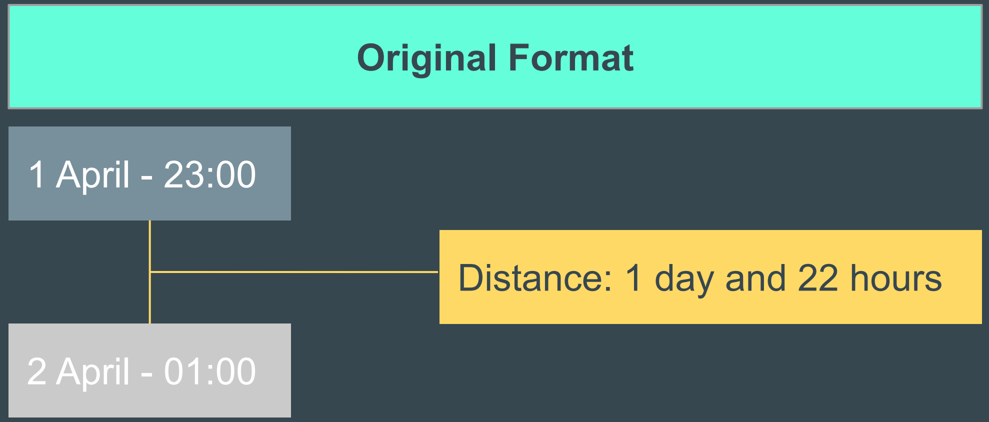
Prediction → Sensor Counts

Data Transformations

All given datasets combined into one

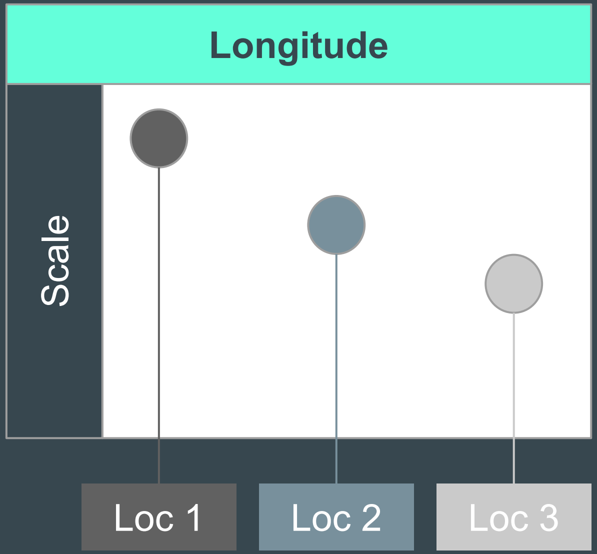
Time

- Problem → Distance between given days and hours unclear
- Solution → Make time circular
  - Separate each month, day, and hour in cos and sin
  - Improved performance significantly



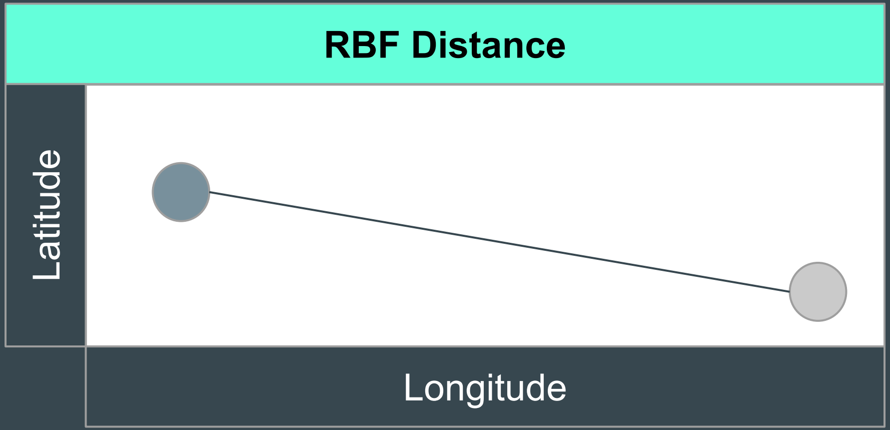
Co-ordinates

- Problem → Placement locations unclear
- Solution → Encode co-ordinates
  - Scale the longitude of all locations
  - Scale the latitude of all locations
- Scaler → Standard Scalar
  - Assumes normal distribution
  - Small performance improvement
  - Adapt at handeling outliers



Distance Stations to Sensor

- Problem → Distance from each sensor to all stations unclear
- Solution → RBF Kernel
  - Euclidean distance longitude & latitude sensor & station
  - Station lowest distance → Highest influence
  - Small performance improvement



Prediction Models

Random Forest

What[1]

- Builds group of weak learners to form strong learner
- Each learner works with subset features → Reduces model complexity
- Prediction → Average prediction all forests

Advantage

- Good performance
- Simplicity in hyperparameter Tuning

[1]Ho, T. K. (1995). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition, volume 1, pages 278–282. IEEE.

XGBoost

What[2]

- Gradient Boosting
- Scalability

Advantage

- High Performance
- Missing Values

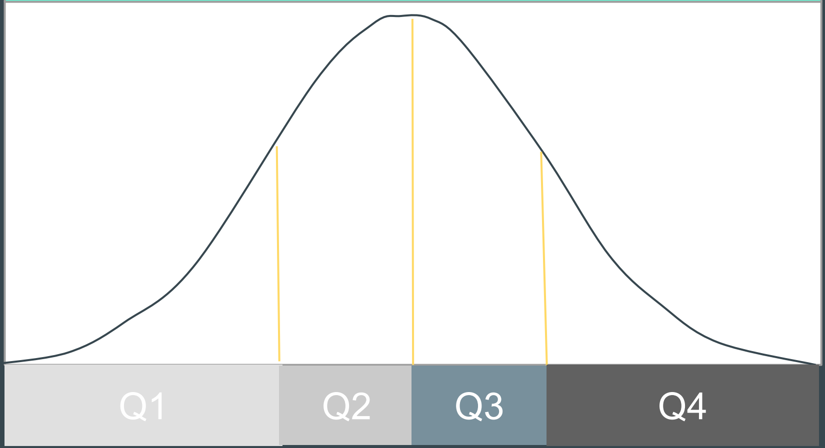
[2] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794. ACM.

Outcome Forms Prediction

Regression	
Prediction	Sensor Crowdedness Counts

Classification	
Prediction	Quartile
Level 1	0% - 25%
Level 2	25% - 50%
Level 3	50% - 75%
Level 4	75% - 100%

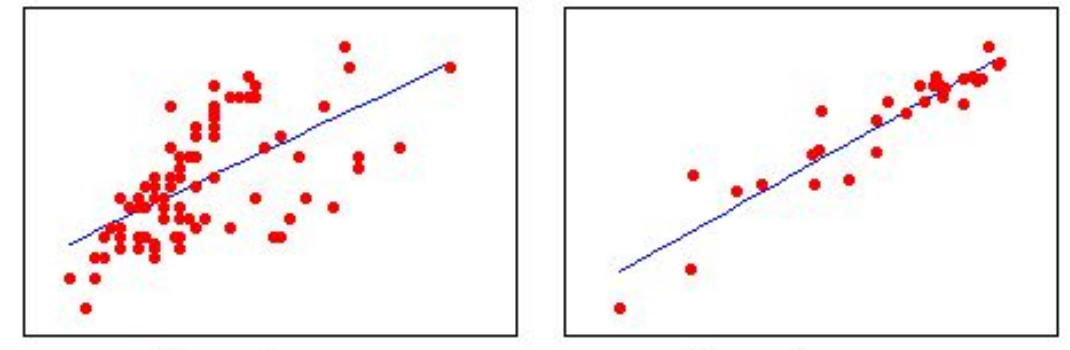
Sensor Crowdedness Counts



Evaluation Metrics

Regression

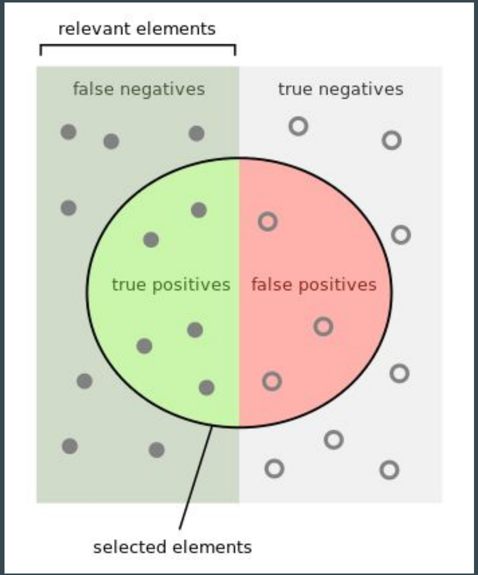
- R^2 → Proportion variance predictable from input variables
- RMSE → Error predicted and true values



Evaluation Metrics

Classification

- Accuracy → Proportion correctly labelled
- Precision → Per class, proportion correctly classified in class
- Recall → Per class, proportion correctly classified as class
- F1 → Balance Precision & Recall



Prediction Method

Model Construction

Split Dataset

80% - Train | 20% - Evaluation

Hyperparameter Tuning → Random Search

100% - Train

Train models → Cross-Validation

90% - Train | 10% - Test

Generate Predictions

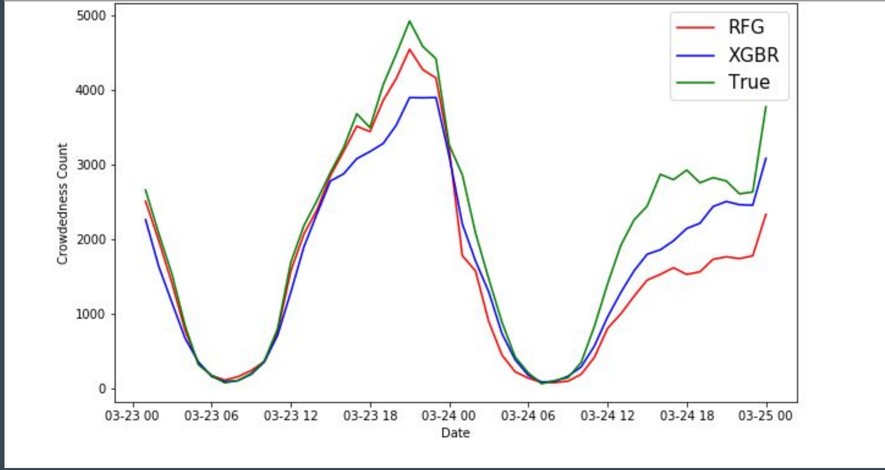
Evaluation

Results

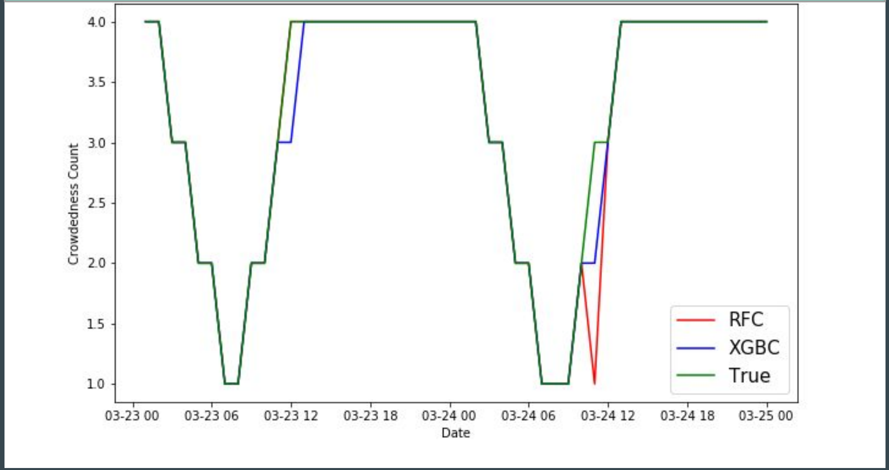
Regression		
Model	**R2**	**RMSE**
Baseline	57.7%	654.1
Random Forest	83.3%	411.27
XGBoost	*85.2%*	*387.28*

Classification			
Model	**Accuracy**	**Precision**	**Recall**
Baseline	24.1%	24.1%	25%
Random Forest	84.4%	84.4%	84.4%
XGBoost	*85.8%*	*85.8%*	*85.8%*

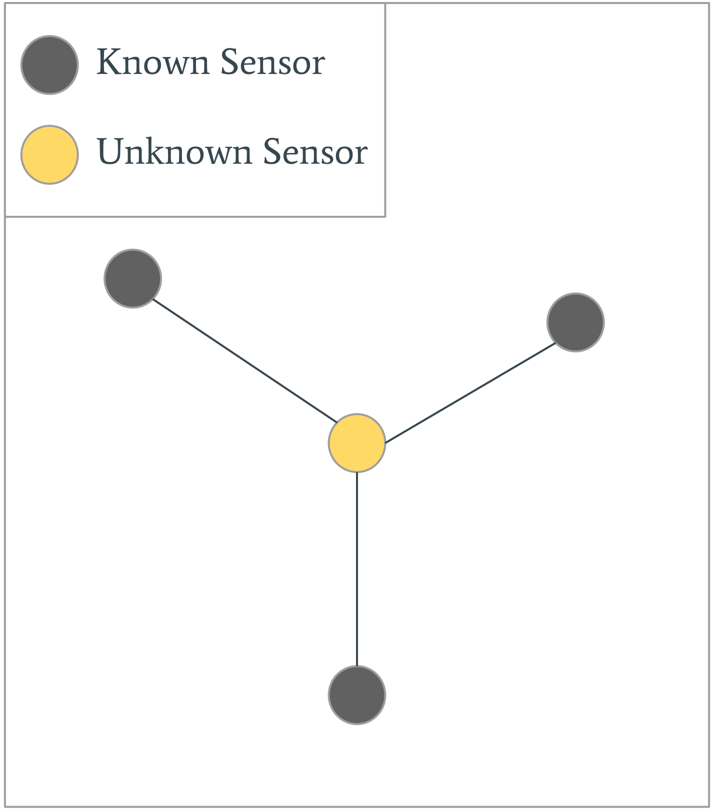
Regression



Classification



Generalization Method



Model Construction

Split Dataset

Train → 6 sensors | Evaluation → 1 sensors

Hyperparameter Tuning → Random Search

100% - Train

Train models → Cross-Validation

90% - Train | 10% - Test

Generate Predictions

Evaluation

Results

Regression		
Model	**R2**	**RMSE**
Baseline	58.3%	656
Random Forest	84.4%	401
XGBoost	*85.5%*	*386.1*

Classification			
Model	**Accuracy**	**Precision**	**Recall**
Baseline	24.1%	24.1%	25%
Random Forest	*84.2%*	*84.2%*	*84.2%*
XGBoost	*84.2%*	*84.2%*	*84.2%*

Wrap up

Discussion

Limitations

- Spatial Dimension not used
- Sensor data affected performance

Recommendations

- Sensor data
- Real-time predictions

Conclusion

- Public Transport data used to predict crowdedness
- Prediction & Generalization returned effective results
- Overall → XGBoost superior

Thank you for your attention

Don de Lange

Msc Data Science