A Comparison of Linear Regression Model and Random Forest Model for Misdemeanour Prediction in Manhattan

XINRAN WANG	s6036163	GFM
MAX KORIR	s6035760	NRM
QIAO REN	s6036341	GFM
UTSAV SONI	s6038387	GFM

Introduction

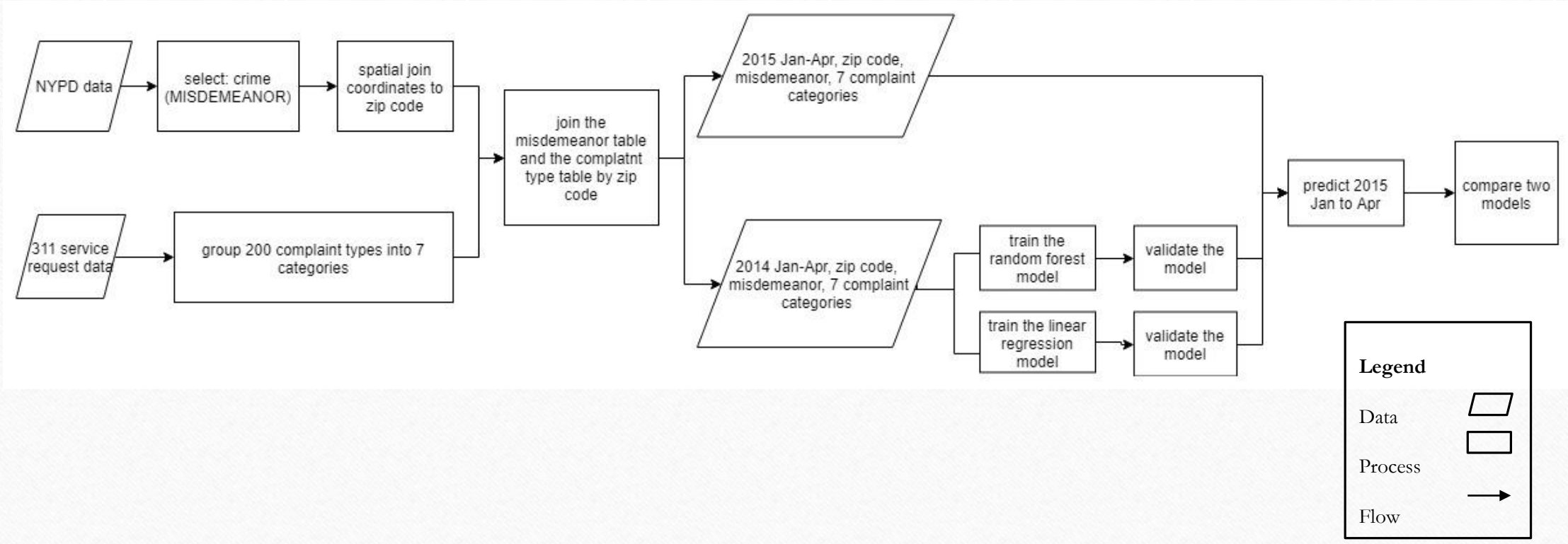
Background information:

Objectives:

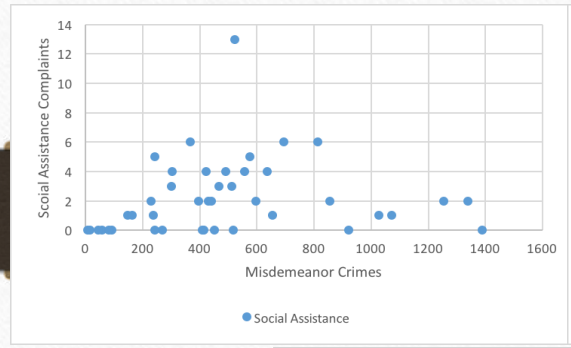
Finding the relationships between complaints and crimes instances in space and time in Manhattan

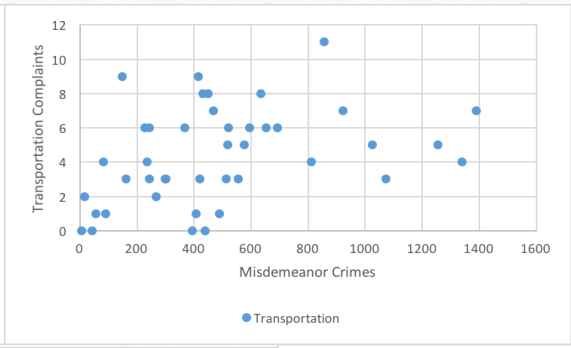
Developing a model to predict the most probable region of crime growth for the year 2015

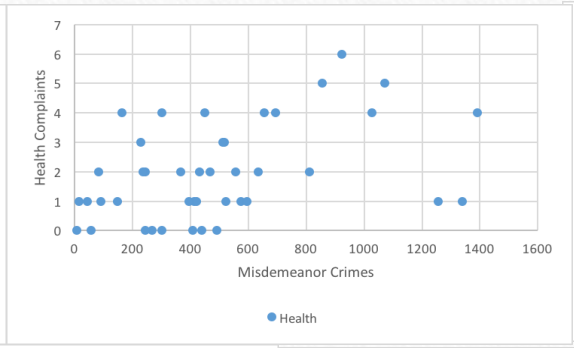
Method

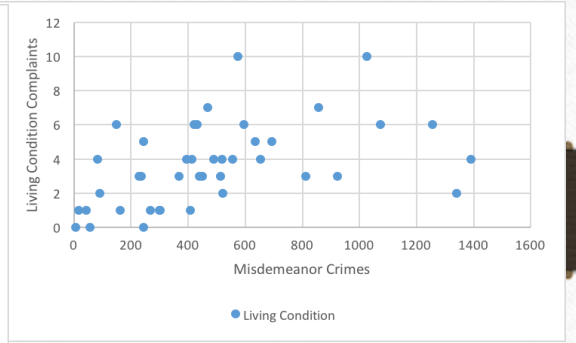

NYPD data
select: crime (MISDEMEANOR)
spatial join coordinates to zip code
311 service request data
group 200 complaint types into 7 categories
join the misdemeanor table and the complatnt type table by zip code
2015 Jan-Apr, zip code, misdemeanor, 7 complaint categories
2014 Jan-Apr, zip code, misdemeanor, 7 complaint categories
train the random forest model
train the linear regression model
validate the model
validate the model
predict 2015 Jan to Apr
compare two models
Legend
Data
Process
Flow

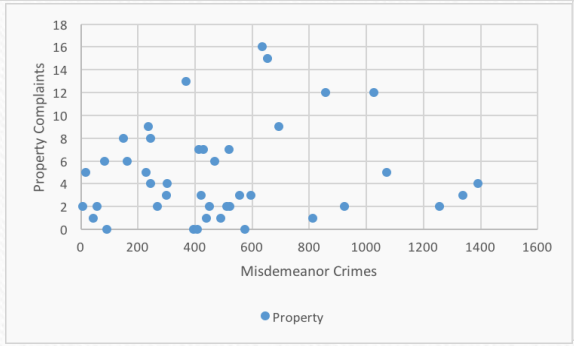
Data Processing - correlation

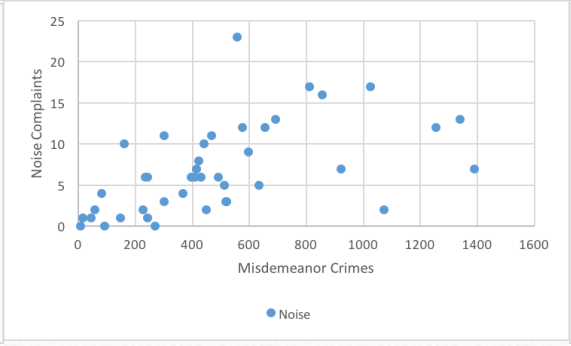


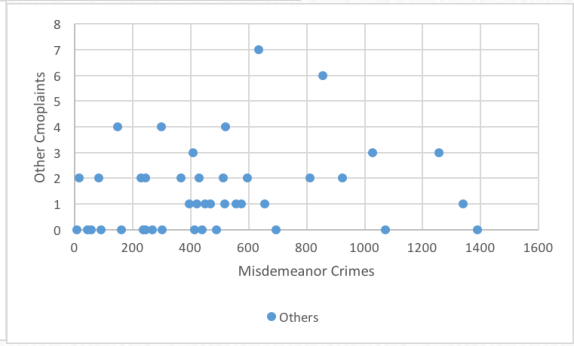










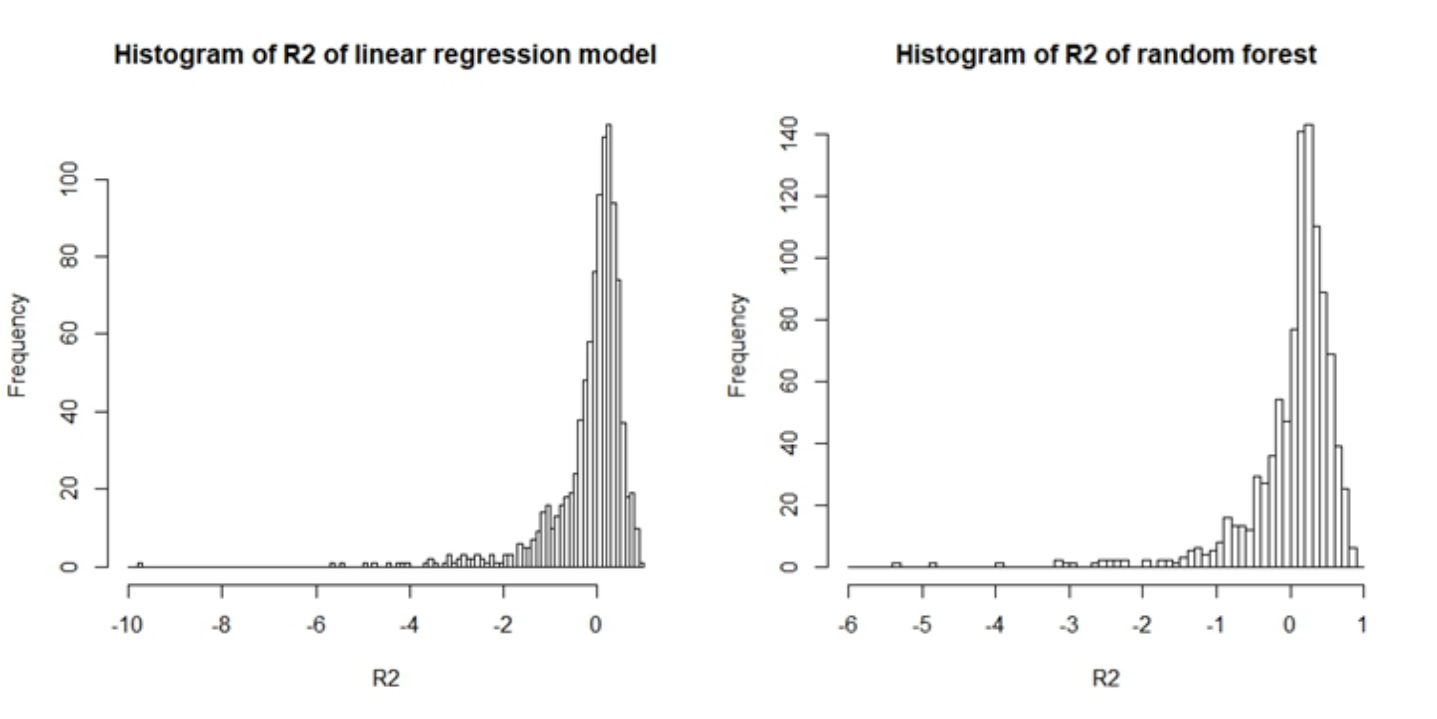


Data Processing - correlation

Health	0.438006653
Living Condition	0.46201815
Noise	0.541775903
Others	0.180127613
Property	0.109604177
Social Assistance	0.12684125
Transportation	0.361190061

Result – Training and Validation

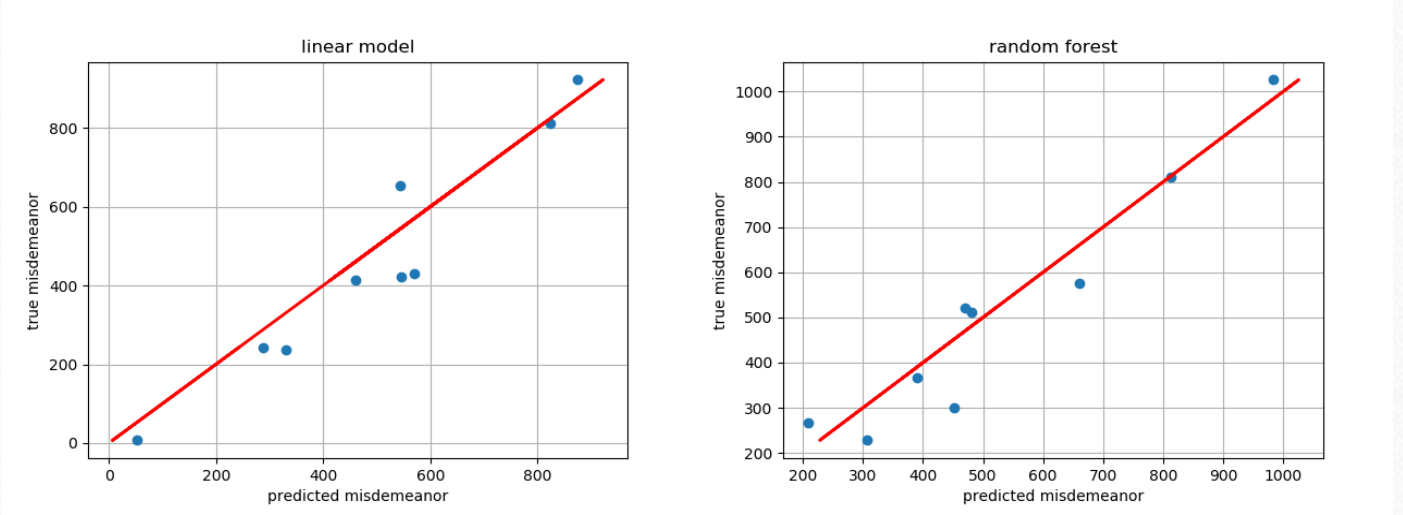
R2 value after running for 1000 times



Result – Training and Validation

Select the best model.

Predicted misdemeanor in 2014 against the observed misdemeanor in 2014

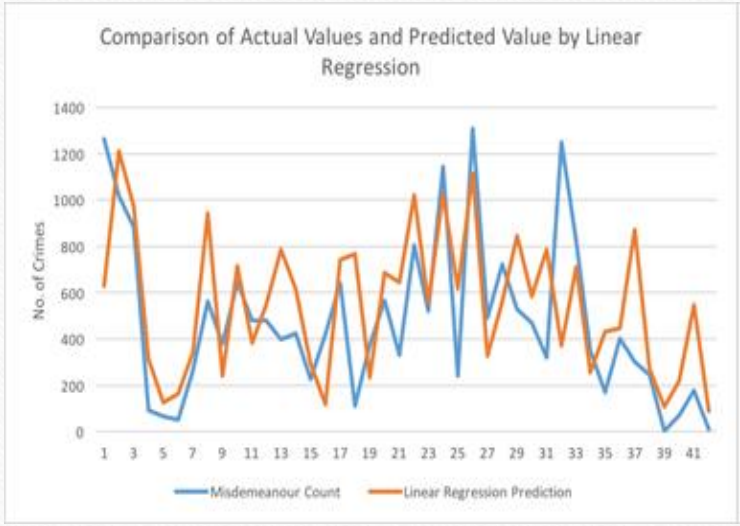


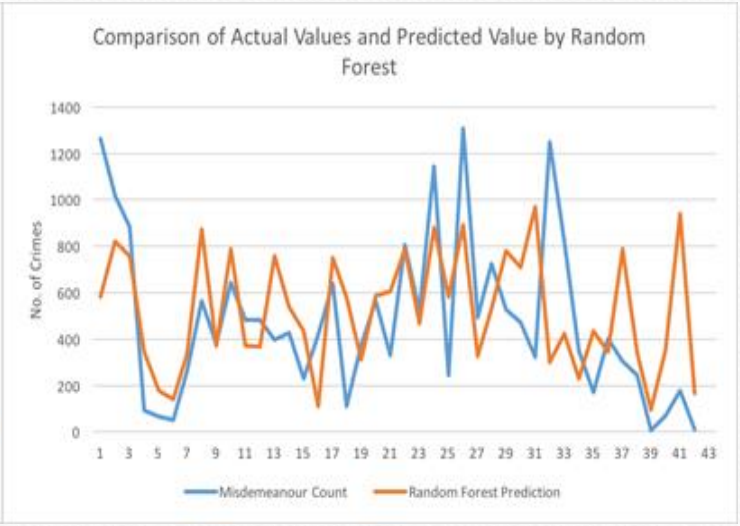
When R2=0.9284

When R2=0.9079

Result - Prediction

Comparison between predicted misdemeanour value and observed misdemeanour value in 2015





Discussion

- In the whole process, we have reasonable conditions and a logical model, which helps to find the most relevant complaint type for crime prediction.
- Some of the limitations in this study are:

i. The size of the dataset

ii. Complaint types aggregation

Conclusion

- The random forest model has a better performance and it is more stable than the linear regression model
- Complaints are correlated with misdemeanour cases at different levels
- This study could be used to predict the trend of misdemeanour frequency in each zone.

Thank you!