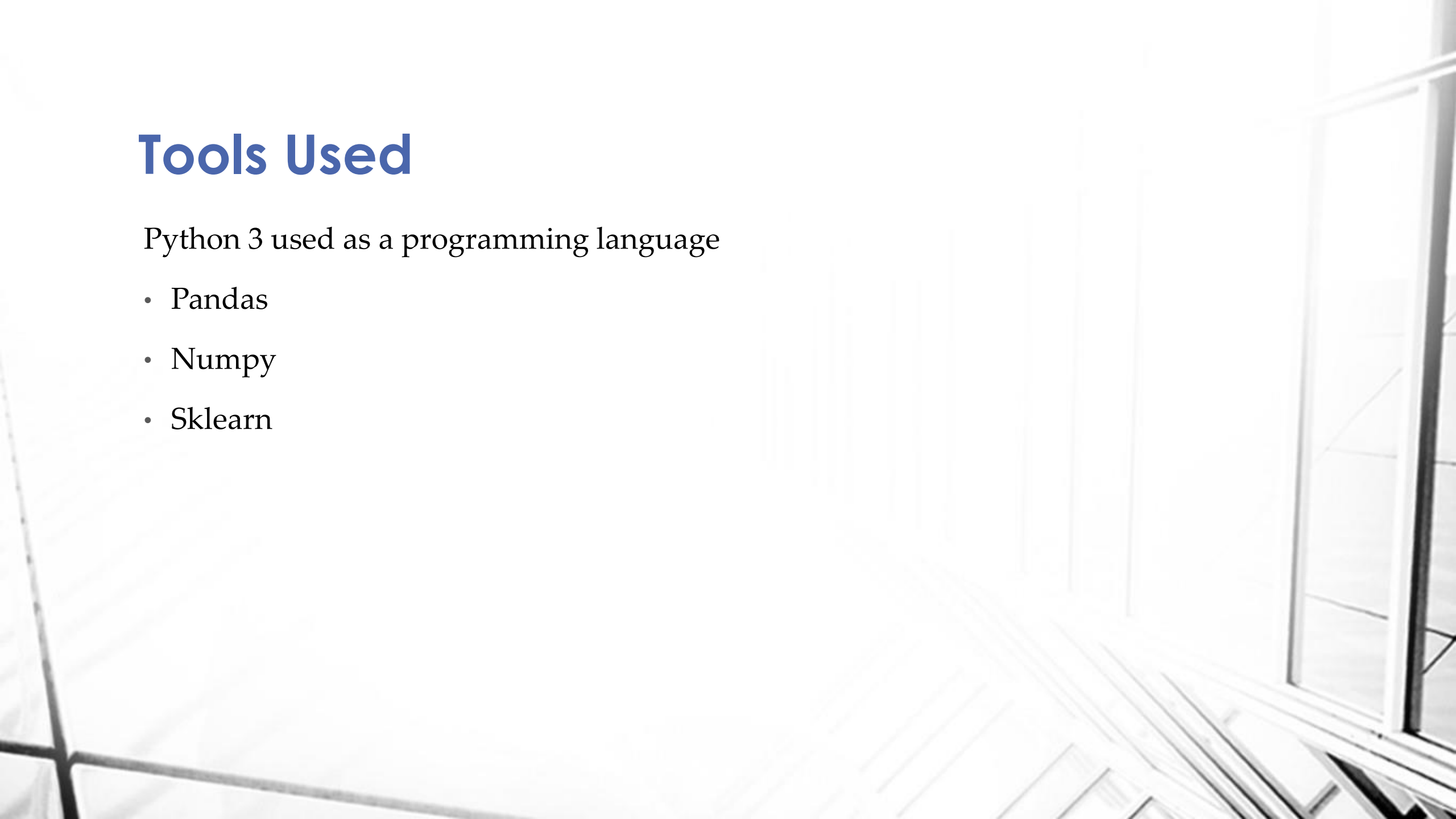
Citi Hackathon-Citi Campus Innovation

Presentation-
Naver Solutions

Tools Used

Python 3 used as a programming language

- Pandas
- Numpy
- Sklearn



Missing Data

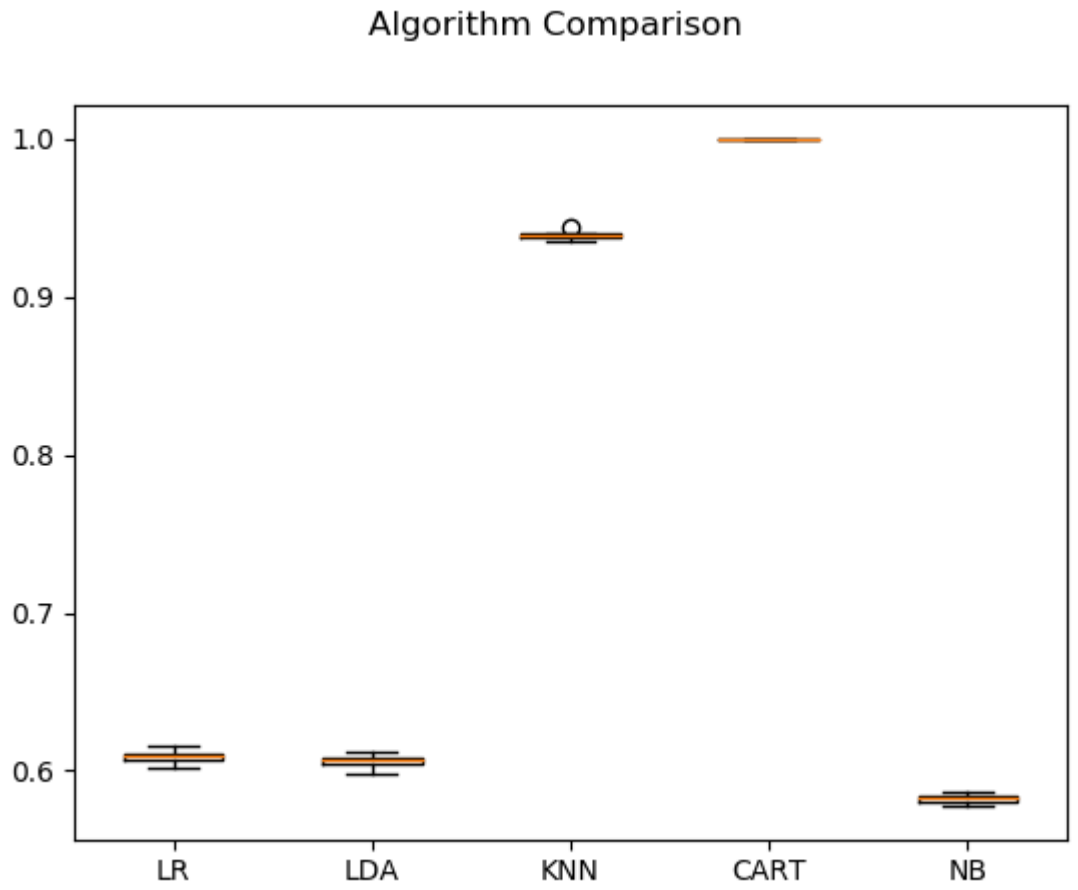
- I observed that certain rows had missing data, so the plan was to use the rows with all the data(i.e. SCORE, BORO ,VIOLATION CODE etc.) for training purpose.
- For training, whenever the values in the rows (SCORE mainly) were missing, we found that “Not Applicable ” CRITICAL FLAG was given in more than ~80% of cases. So we omitted them and predicted label as “Not Applicable” whenever the test row has NaN value in SCORE.

Feature Engineering

- By inspection we found that the following attributes were not having significant effect on the CRITICAL FLAG, which were removed from training:
  - UIDX, CAMIS, PHONE, DBA, BUILDING, ZIPCODE, INSPECTION DATE, GRADE DATE, RECORD DATE.
- Attributes like VIOLATION DESCRIPTION was just the description of VIOLATION CODE. So VIOLATION DESCRIPTION were removed
- Lastly, we changed all the labels in the columns to particular numeric values (0,1,2…etc).

Model Implemntaion

- We implemented different types of classifier models to predict the labels
- And after comparison we selected CART = Decision Tree Classifier as it gives the best result among all the others.
- LR: 0.608801 (0.003445)
- LDA: 0.605990 (0.003590)
- KNN: 0.695689 (0.003812)
- CART: 1.000000 (0.000000)
- NB: 0.582517 (0.002338)



End Results

- We observe that “Inspection Type” was not affecting the accuracy of the model so we removed it to get better results
- We also used models like Random Forest, SVM and Bagging Classifier but they didn't produce improvement over the past results but further take long time to process
- We got a decent accuracy of 48.08531 in comparison to the previous submission by using Adaboostclassifier.