AI Explorers

Starring:

Leidy **Andres** **Dimitris** **Carlos** **Hesheng**

“...to boldly go where no one has gone before”

BUSINESS PROBLEM

INITIAL BUSINESS PROBLEM PROPOSED

Predict the daily number of interventions for each Fire station in the city of Montreal

FINAL PROBLEM SOLVED

01 Prediction

Predict Fire Risk Level for each house in the city based on distance from fire incidents

02 Prevention

Determine property-level fire risk: likelihood of a given property having a fire incident in a given 6-month time period

EXTERNAL DATA USED

01

House Risk Level - based on distance from fire incident

1. The interventions data from the fire department of the city of Montreal

2. Sample of house's data from Centris

02

Property Fire Predictor - based on four (4) closest incidents from center of borough

1. The interventions data from the fire department of the city of Montreal
2. The 2016 Census for the Agglomeration de Montreal
3. The data about Crime in Montreal
4. Data on the properties' location, size, year built, lot area, etc. from the city of Montreal

DATA ANALYSIS TOOLS & TECHNIQUES

01

House Risk Level - based on distance from fire incident

Tools

- Github
- Python
- Java
- Jupyter Notebook
- Excel

Techniques

- Decision Tree Regression

02

Property Fire Predictor - based on four closest incidents from center of borough

Tools

- Github
- Python
- Alteryx
- Tableau
- Excel

Techniques

- Decision Tree
- Random Forest
- XG Boost

DATA ANALYSIS TOOLS & TECHNIQUES

Google Cloud Platform & Alteryx

Data Exploration	Data Storage	Data pipelines	Build ML models
At very first step we required to understand an initial data trend and patterns that could drive us to a possible need in terms of ML. Tools: Data Studio ALTERYX	We rapidly noticed a huge need in terms of storage for all the data we were gathering for a broad of purposes, the picked tool must fit with any format of data. Tools: BigQuery Cloud Storage	In order to cover all the different perspectives we had on the ML model it was required to explore external data that had to be manipulated to fit within the project Tools: Cloud Datalab Cloud Dataflow ALTERYX	On the final stage, once all the workflow had been achieved successfully we finally proceeded to design and build the ML models would fulfill the prediction needs. Tools: Cloud Datalab ALTERYX

MODEL 01 - ASSUMPTIONS

01

House Risk Level - based on distance from fire incident

1. Sample Centris data
    - The model considered some Montreal houses only – discarded other types of constructions e.g. schools, commercial buildings, etc.

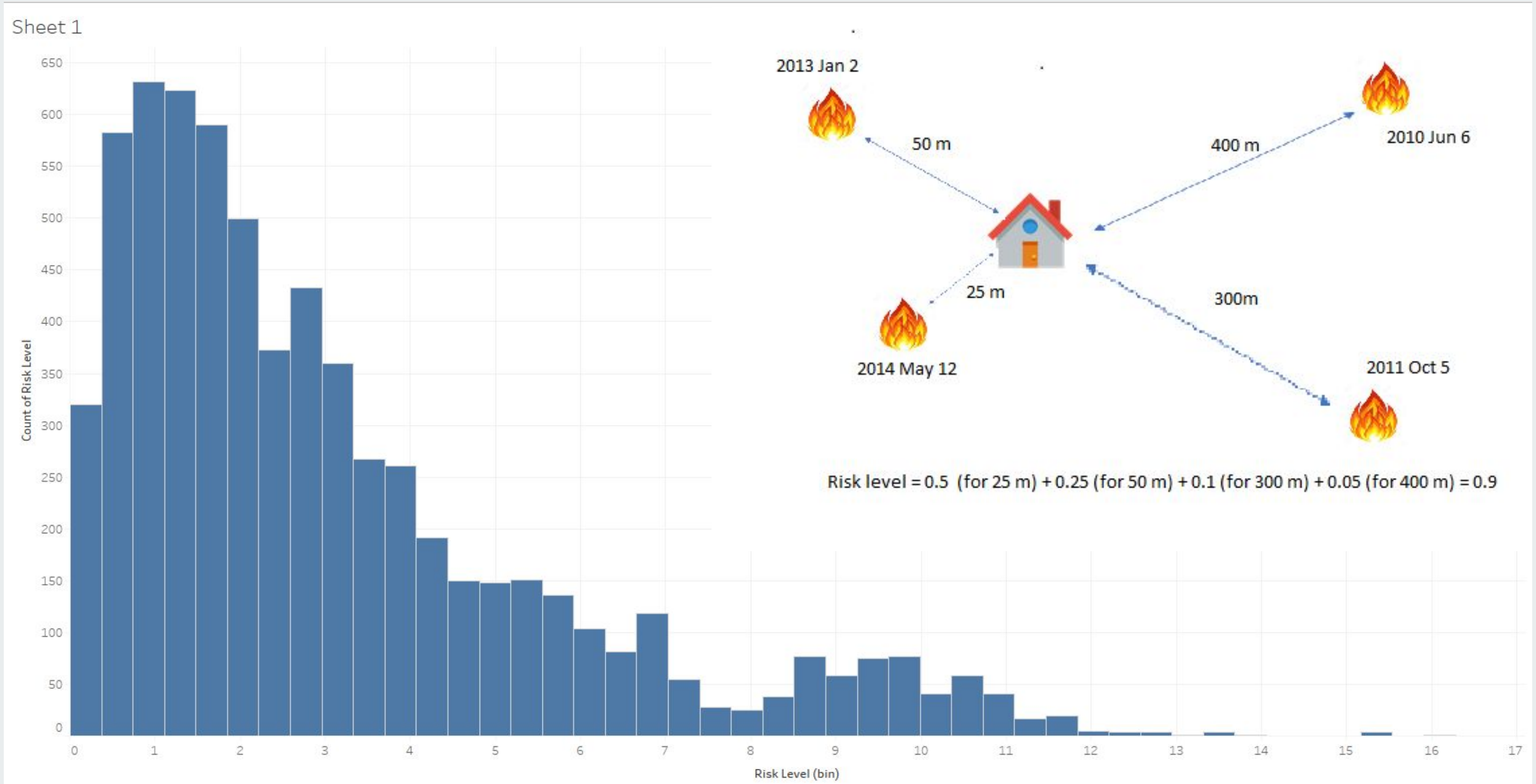
2. Used fire incidents only – discarded other types of incidents e.g. flood, traffic accident, etc.

3. Used house's characteristics only as the features for the model, such as lot size, borough, year built, price, etc. – did not use external additional data

MODEL 01 - ASSUMPTIONS

01

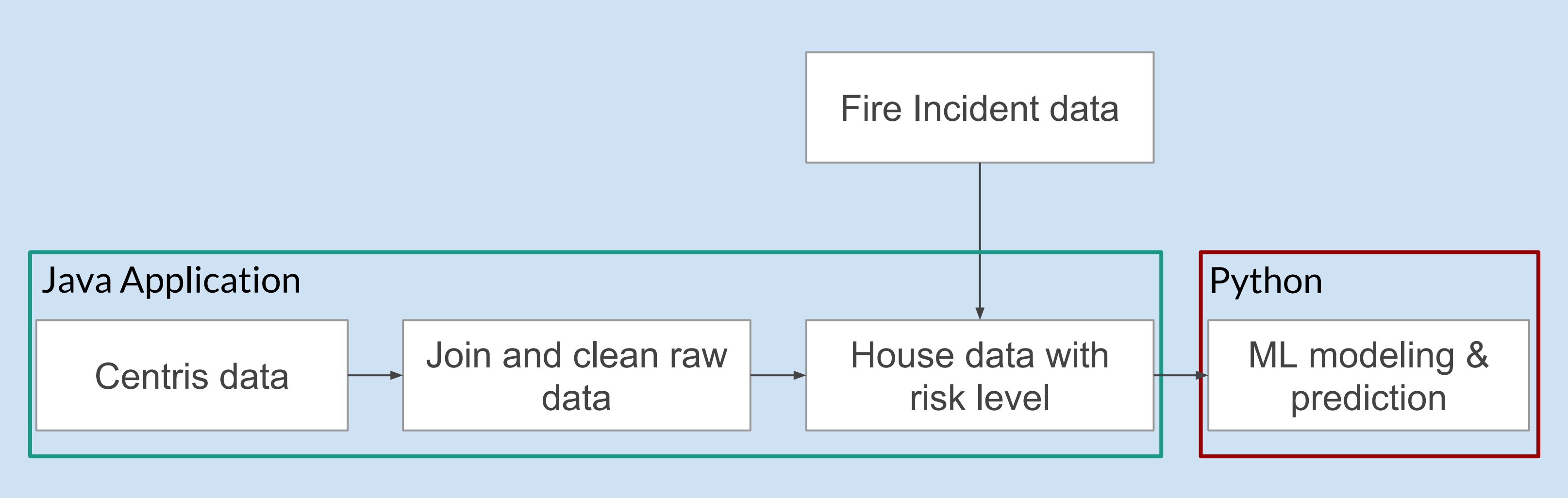
Assign house fire risk level based on distances from past fires



MODEL 01 - DATA PREPROCESSING

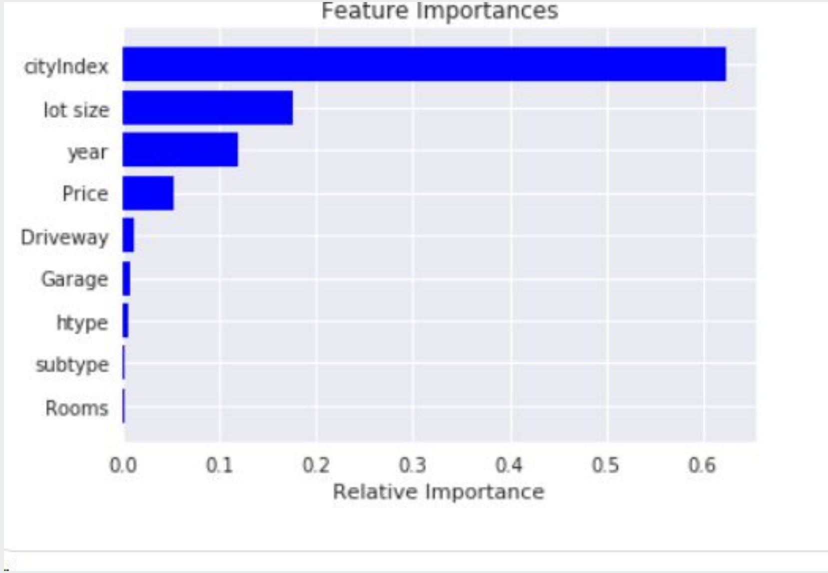
01

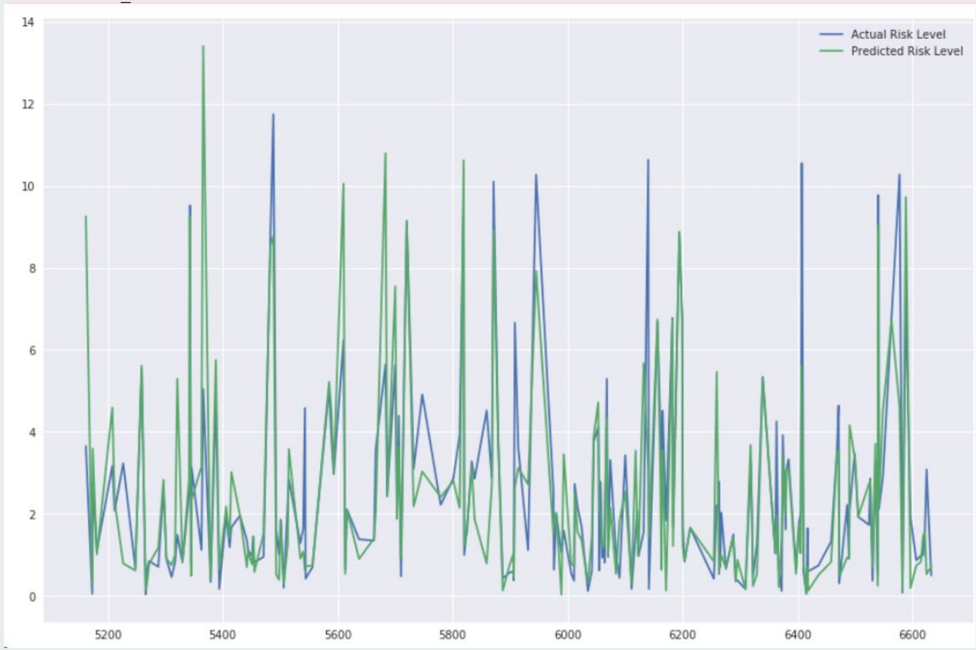
Data Pipeline



MODEL 01 - FINDINGS & RESULTS

Feature Importance





Model statistics summary

Mean Absolute Error: 1.03
Mean Squared Error: 3.26
Root Mean Squared Error: 1.80
R2_score: 0.59

MODEL 01 - FINDINGS & RESULTS

House Risk Level - based on distance from fire incident

The threshold for cutting 5% higher risk level houses is 9.2

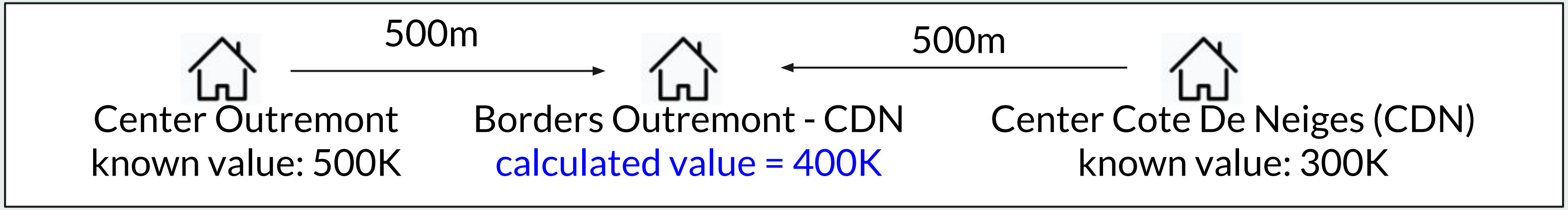
	Predict Not High Risk	Predict High Risk
Actual Not High Risk	781	23
Actual High Risk	25	29

MODEL 02 - ASSUMPTIONS

02

Six Months Fire Predictor Modeling

1. Assumed that the data from Census Canada is smoothly distributed across the boroughs as show here: (*We actually used values from 4 boroughs, but here we show only 2 for simplicity*)



2. The model used 6 months of historical (past) data and looked 6 months into the future

Features not included in the model		**Features**		**Label**
Location - not used as a predictor in the model, only to join data	Date - not used as a predictor, only to create table	**History related features** looking back 6 months (7 features) Sum_Autres_incendies, Sum_Incendie_de_batiments, Sum_Premier_Repondant, Sum_Sans_incendie, Sum_Alarmes_incendies, Sum_False_Alertes_Annulations, Sum_Crime.	Location related features (25 features)	Was there a fire in the NEXT 6 months? (look into the future to respond)

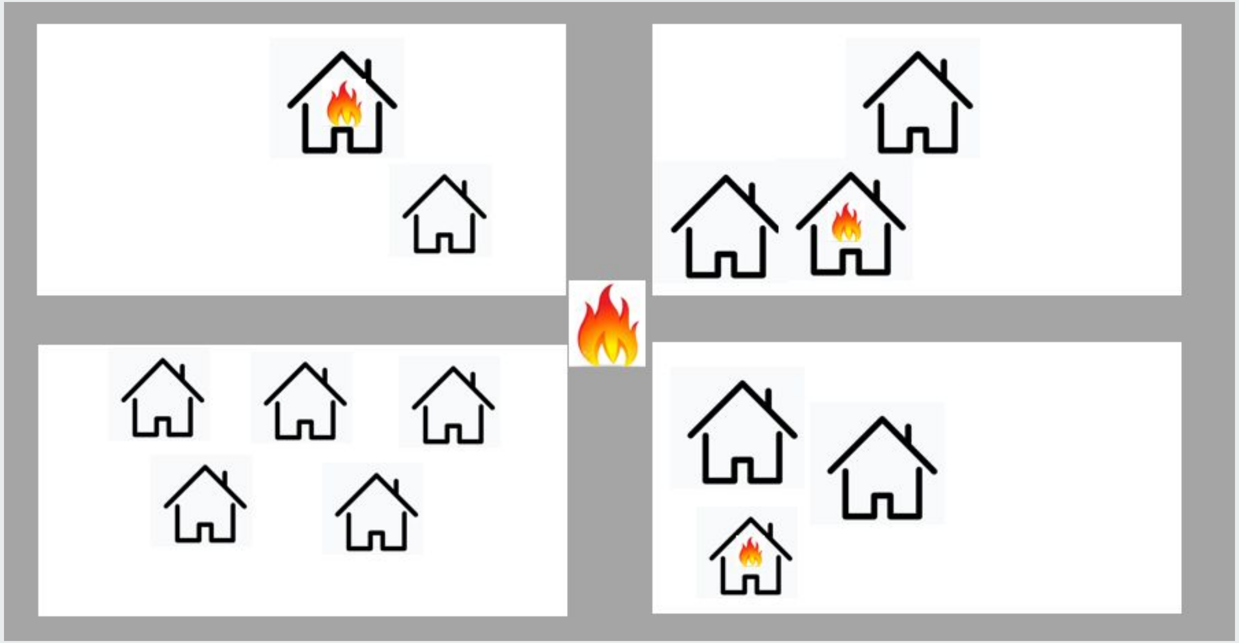
MODEL 02 - ASSUMPTIONS

02

3. Assumed that the street intersection where the incident is projected is the location where the incident occurred and we estimate its features as the average of the properties that project there.

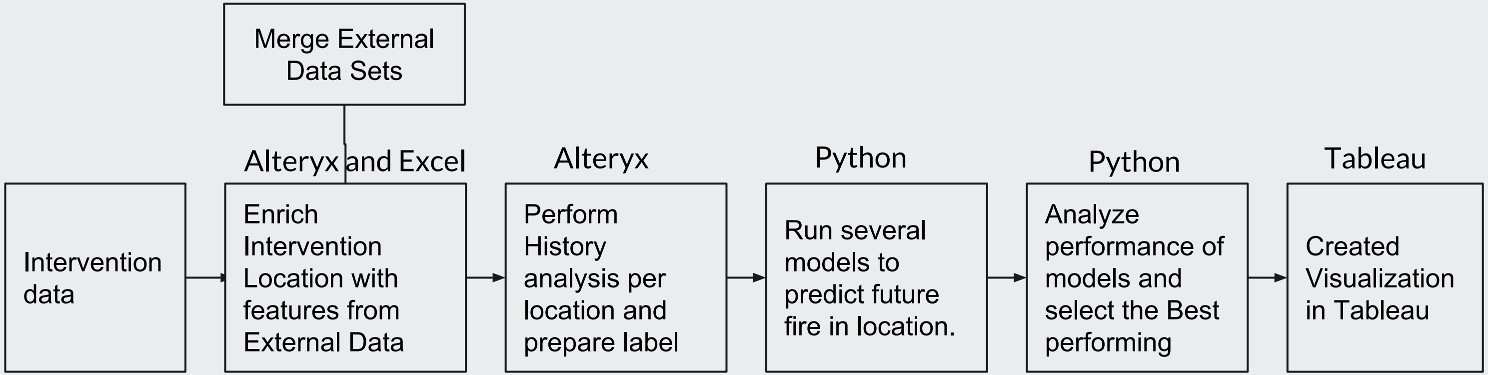
Features used (averaged)

1. Year built
2. Lot size
3. Area of the house
4. Number of floors



MODEL 02 - DATA PREPROCESSING

Data Pipeline



MODEL 02 - FINDINGS & RESULTS

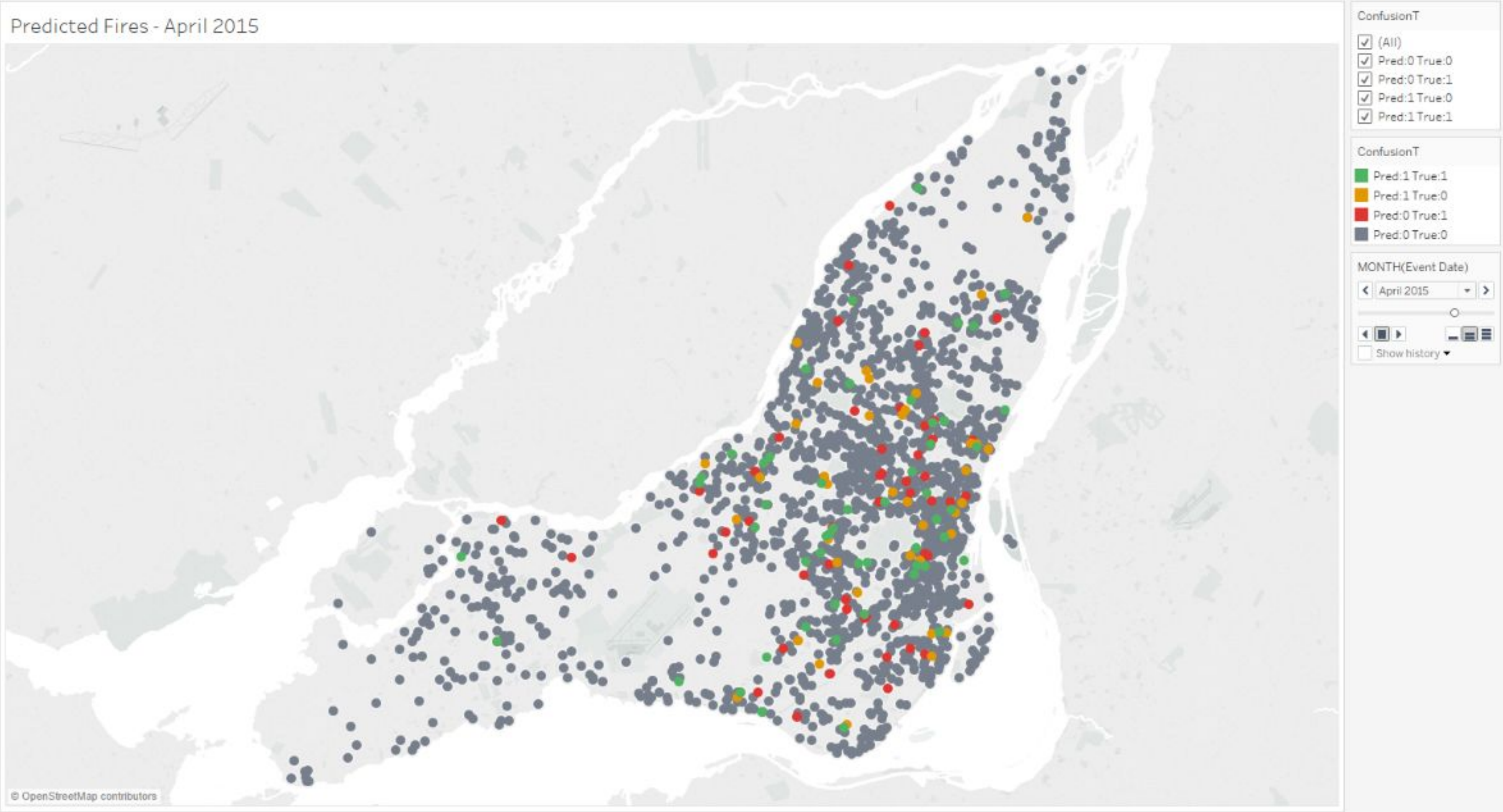
Summary of the three (3) ML models we used - Winner is Random Forest

Model	Recall	Precision	F1_Score	Kappa	AUC
Decision Tree	45.5%	43.9%	44.7%	40.3%	0.72
Random Forest	**44.6%**	**61.7%**	**51.8%**	**48.6%**	**0.90**
XG Boost	12.9%	81.4%	22.3%	20.7%	0.90

Confusion Table 34% of total data		**Predicted**	
		No Fire	**Fire**
Actual	**No Fire**	**181,173**	**4,045**
	Fire	**8,105**	**6,524**

- XG Boost misses fires for the sake of better prediction accuracy. This is why this model was not selected as the winner.

MODEL 02 - FINDINGS & RESULTS



PROJECT MANAGEMENT

Tools used to develop and manage the project

1. Project Management: Trello, Team Charter
2. Presentation: Prezi, Google Slides
3. Communication & Collaboration: Slack, Email & Tasks on Trello board, Hangouts
4. Data Visualization: Tableau, Python, Data Studio
5. Data Preparation: Excel, Data Lab, Big Query, Python, Alteryx, Java
6. ML: Python, Alteryx, Data Lab, Jupyter Notebook, Colaboratory
7. Work Methodology: Kanban on Trello
8. Github
9. Google Drive, Google Suite: Docs, Sheets

Project Deliverables

- Two PDF reports: House Risk Level Modeling & Six Months Fire Predictor Modeling
- ML Models
- Project Documentation & related work e.g. Alteryx & Python workflows

How will the project be archived?

- Github & Google Drive

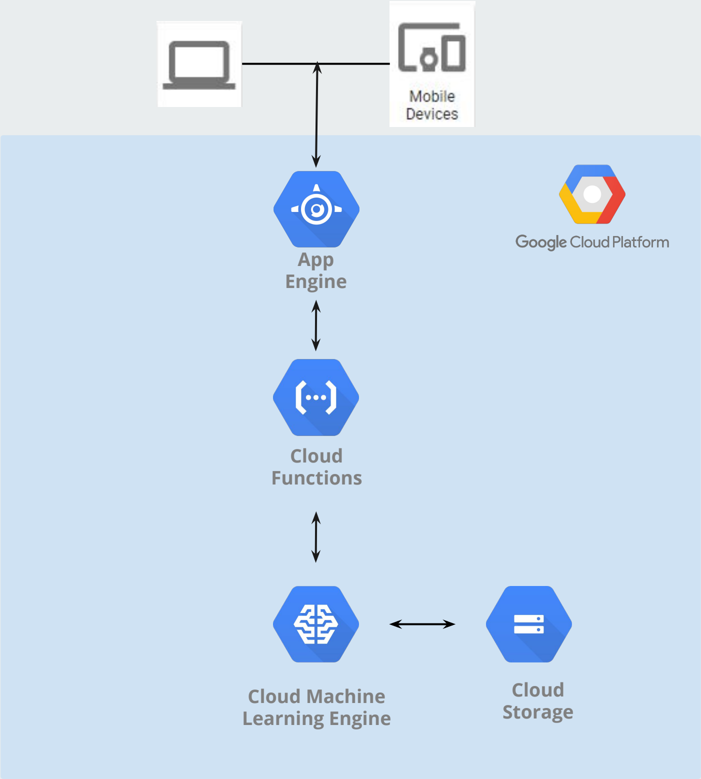
CHALLENGES & CONSTRAINTS

Challenges the team faced

1. Not enough data available to train and test the model
2. Quality of data: obfuscated data without actual property address
3. Hard to meet in person as all team members were busy during business hours
4. Difficulties in agreeing where we wanted to go and what we wanted to predict
5. Difficulties in finding additional external data to leverage the quality of data
6. We found the project proposed too generic and hypothetical, therefore subject to open interpretations
7. Difficulties setting up GCP platform - new tech. learning curve
8. Not enough time in class to discuss the project
9. Not enough time to present and explore work done during the presentations

ML DEPLOYMENT FLOW / DEMO

ML model deployment pipeline



Sample code to export model

Plain Decision Tree Model

```
#Decision Tree
from sklearn.externals import joblib
from sklearn.tree import DecisionTreeClassifier
dtree = DecisionTreeClassifier(max_depth=60,random_state=167)
dtree.fit(X_train,y_train)
#Export model
joblib.dump(dtree,'model.joblib')
!gsutil cp model.joblib gs://projectcsv/model.joblib
```

Create the input file for our model

```
df = X_test.head(5)
df.to_csv('input',index=False,encoding='utf-8')
!gsutil cp input gs://projectcsv/input
```

```
Copying file://input [Content-Type=application/octet-stream]...
/ [1 files][  2.2 KiB/  2.2 KiB]
Operation completed over 1 objects/2.2 KiB.
```

Output sample

https://us-central1-mcgillcapstone.cloudfunctions.net/launch_prediction

CONCLUSIONS & FUTURE WORK

- **the first one** predicting fire risk level of houses based on distance from incidents.

- **the second** predicting the likelihood of fire in a specific intersection in the next 6-month period based on the location history and characteristics.

Finally we believe that this work, along with the other deliverables produced, if leveraged with more and real data, would prove to be useful for the Fire Department as a tool to improve public safety, resources planning and allocation, and detect and combat fires in the city of Montreal.

AI Explorers

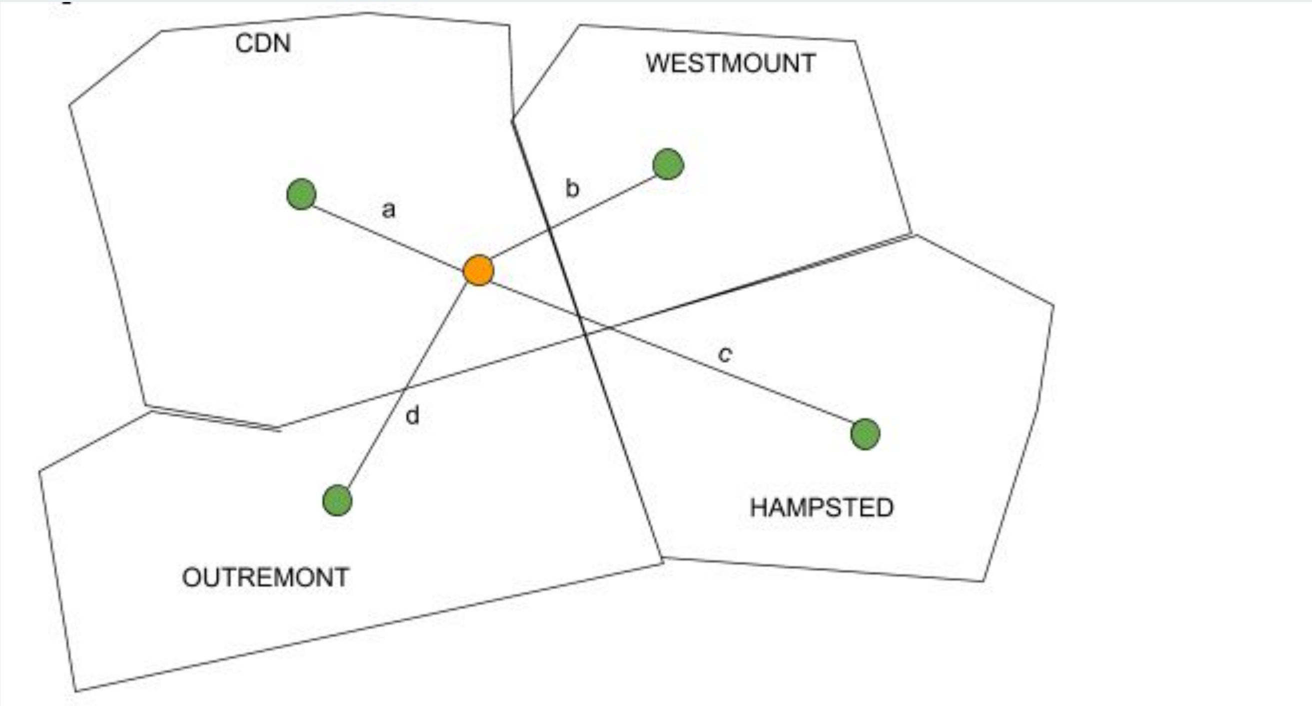
THANK YOU!

APPENDIX

MODEL 02 - ASSUMPTIONS

02

Calculate the center of the borough



MODEL 02 - FINDINGS & RESULTS

02

Feature Importance

0	Sum_Autres_incendies	5.627817
1	Sum_Incendie_de_batiments	4.416537
2	Sum_Premier_Repondant	23.562459
3	Sum_Sans_incendie	14.230483
4	Sum_Crime	0.223628
5	Sum_Alarmes_incendies	12.848393
6	Sum_False_Alertes_Annulations	0.466472
7	Couples_No_Children	1.352183
8	CouplesWithChildren	1.376694
9	SizeOF House	1.478065
10	Detached_House	1.527576
11	Appartment_FiveFloors	1.497550
12	OtherType	1.347506
13	Semi_detached	1.426204
14	TownHouse	1.550613
15	Duplex	1.474393
16	Appartment_less_5Florrs	1.332883

17	OtherDtached	1.456652
18	MObileHome	1.499862
19	1_4Rooms	1.328204
20	5_rooms	1.365861
21	6_rooms	1.373940
22	7_rooms	1.387381
23	8_RoomsOrmore	1.461079
24	Average_Rooms	1.424583
25	Simple_Maintenance	1.289666
26	Major_repairs	1.351905
27	Average_House_Value	1.445947
28	AverageHouseholdIncome2015	1.485375
29	Avg_flors	1.853714
30	Avg_YearBuilt	1.669168
31	Avg_LandArea	1.852248
32	Avg_HomeArea	2.014959

TEAM CHARTER

McGill | School of Continuing Studies | École d'éducation permanente

Team charter
Data Science Capstone Project

1. What is the team name?

AIExplorers

2. What are the team goals and values?

goals : deliver values to the customers

values: transparency, team collaboration.

3. How will the team communicate?

slacker: /whatsup.

TRELLO - KANBAN BOARD