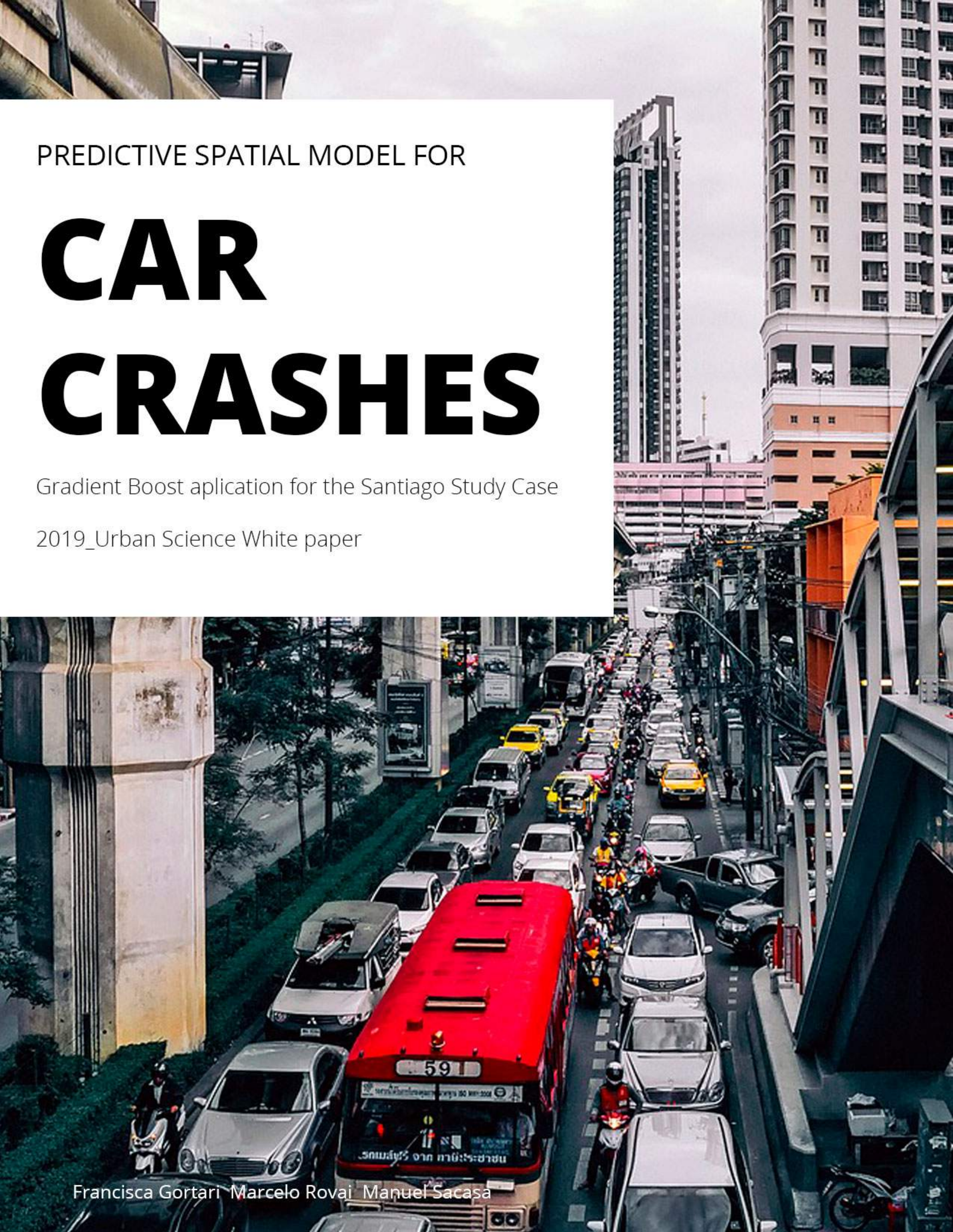
PREDICTIVE SPATIAL MODEL FOR

CAR CRASHES

Gradient Boost aplication for the Santiago Study Case

2019_Urban Science White paper



Francisca Gortari Marcelo Rovai Manuel Sacasa

CONTENT

A BRIEF INTRODUCTION

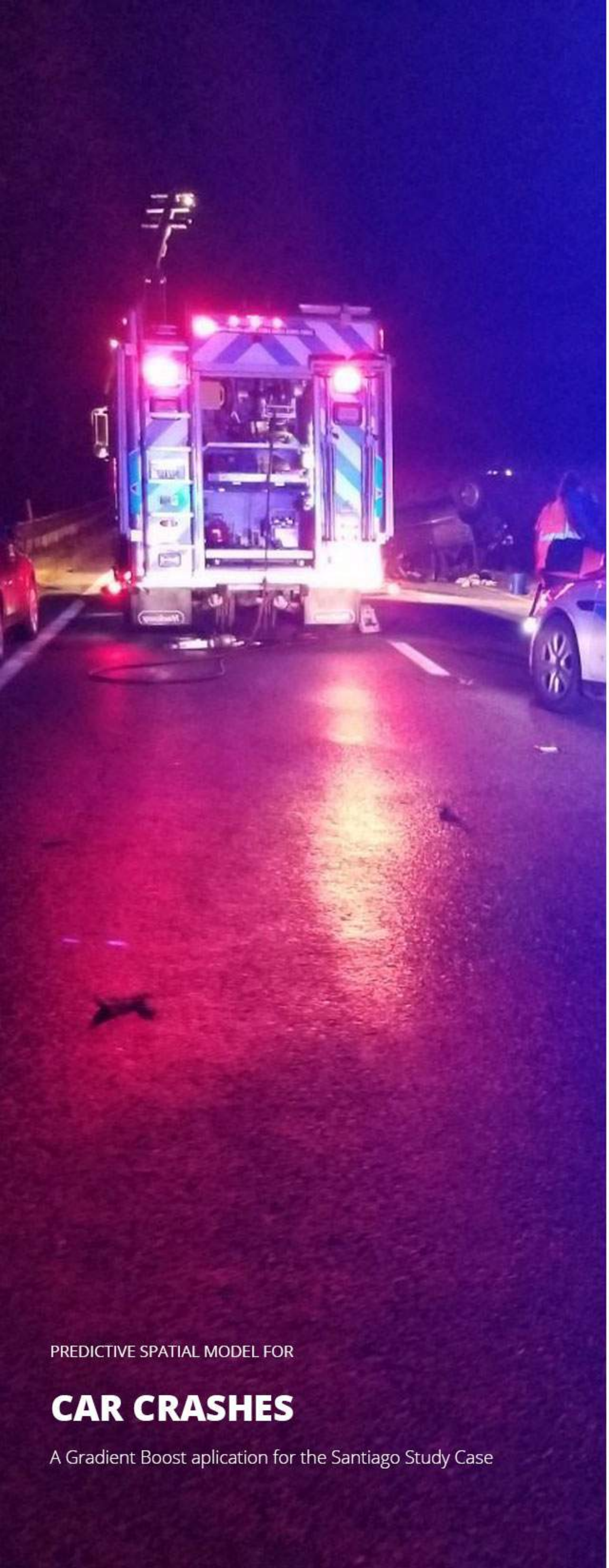
Car crash prediction using Gradient Boost algorithm, CONACET and OPEN STREET MAP datasets.

A car crash predictive model is a highly valuable tool

A car crash predictive model is a highly valuable tool for inhabitants, insurance business, logistic transport and obviously urban data science. This white paper is an urban science wrap up of the process of designing, testing and building a machine learning spatial model to predict car crash risk scoring on a specific grid zone.

Santiago (Chile's Capital and its main inhabitant density city) was the geographic territory selected for this primary model. As the dataset, CONACET 2013-2018 Santiago's car crash data was used. CONACET is an official dataset with descriptive features from the crashes. Some of these features are latitude, longitude, type, severity, number of people injured, level of injury, etc. This crash variables or features, were named as "dynamic", once is directly related to the accident.

Another important source of data was the OPEN STREET MAP "Chile-Latest-free" dataset, with geo referenced points: places, points of interest, traffic, transport, roads and intersections (created from roads). OSM is a non-official dataset with descriptive features from the urban environment. This environment variables were named as "Static", once they are independent from the crash events and for the purpose of this work, assumed immutable over the years.



PREDICTIVE SPATIAL MODEL FOR

CAR CRASHES

A Gradient Boost aplication for the Santiago Study Case

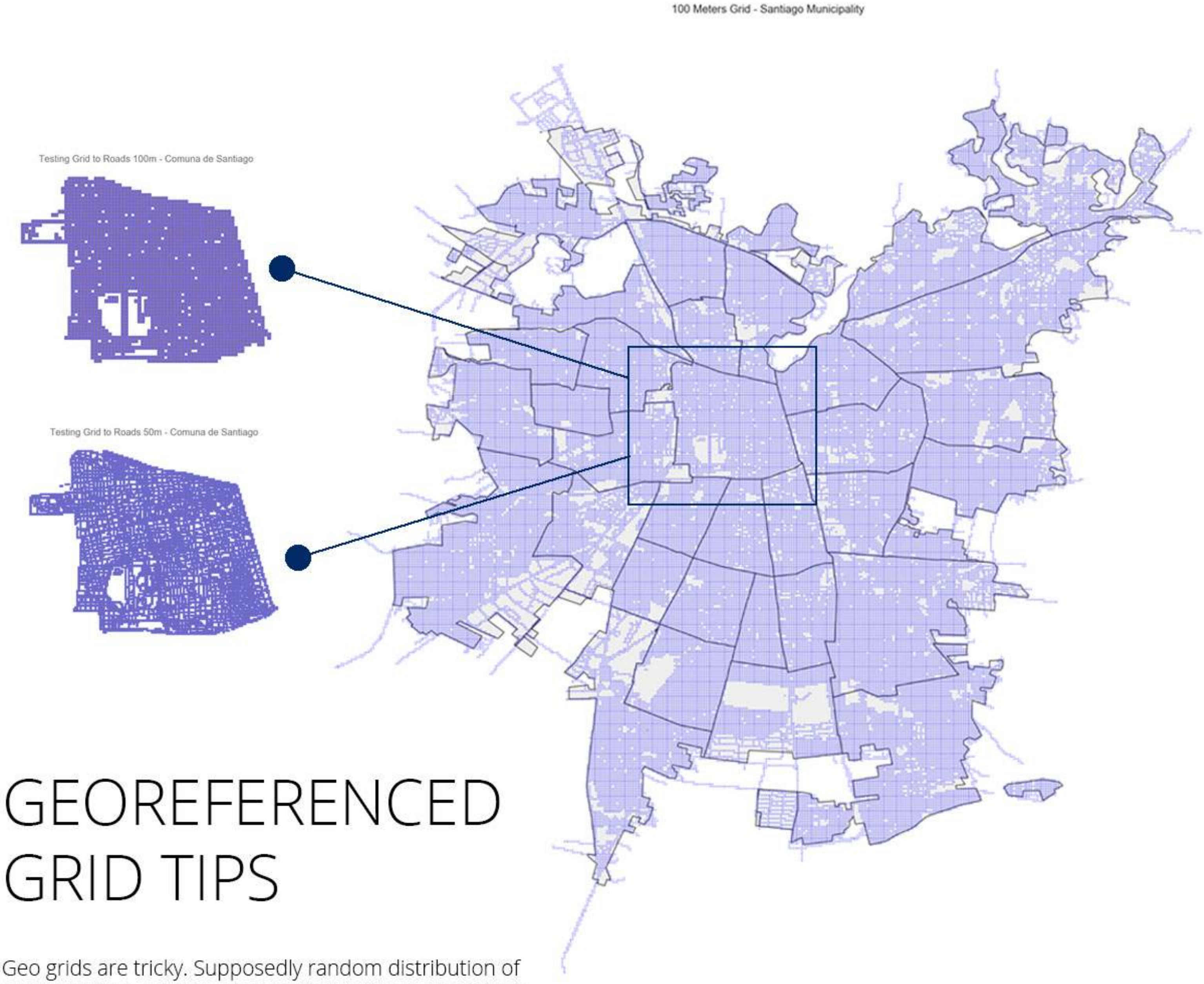
An urban science study case

The study looks for a way to compare geographic zones, were attributes and crashes are supposed to be randomly distributed

The Santiago risk-scoring prediction was deployed based on a discrete grid merging static and dynamic variables. The study looks for a way to compare geographic zones were attributes and crashes are supposed to be randomly distributed. The only way to compare it, to build a discrete geographic unit or a grid.

This grid was enriched with static and dynamic features. The enrichment of each cell grid was developed joining each grid cell (Polygon) with a point (Inside the polygon), line string (Inside polygon) o polygon (Same face or entire inside the cell). Both features and grid may be built with degree space latitude – longitude metric.

When the grid is entirely enriched with all static and dynamic features, it is saved as a csv file to be fed in the next stage, the machine learning preparation process. On this stage, the dataset will be split into training and test sub-datasets to fit the ML models. ML Models as Random Forest (RM), GBT (Gradient Boosting Trees) and a scalable end to-end tree boosting system (XGBoost) were used on this work.

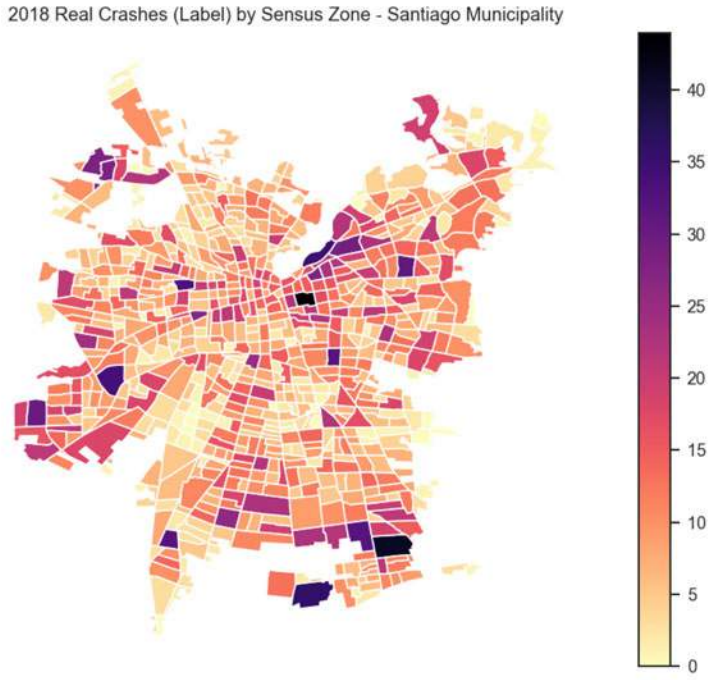

100 Meters Grid - Santiago Municipality
Testing Grid to Roads 100m - Comuna de Santiago
Testing Grid to Roads 50m - Comuna de Santiago

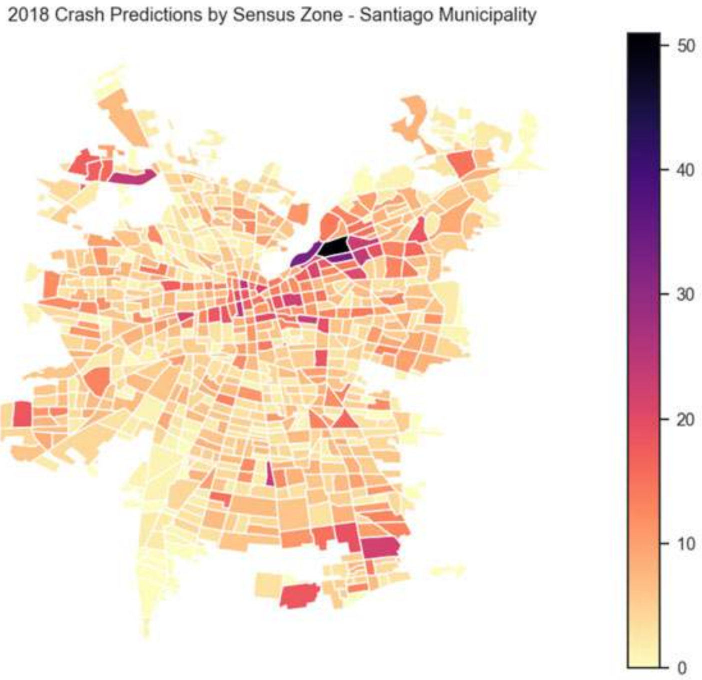
GEOREFERENCED GRID TIPS

Geo grids are tricky. Supposedly random distribution of crashes mixed up with a giant urban surface, resulting in an unbalanced dataset that must be controlled from the beginning. If you don't want to suffer with an unbalanced machine learning train/test process these are the tips:

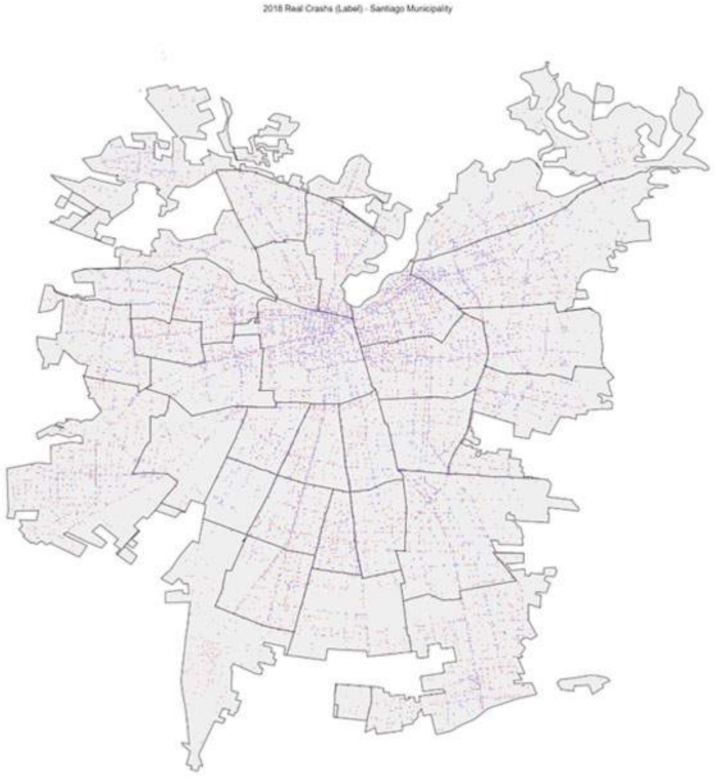
1. Grid cell dimensions. The dimension (or length) of an individual cell will determine the total number of cells necessary to cover the entire city. With small cell dimensions the grid would have a big amount of cell so your final dataset will have a large amount of sample with Crash equals to "0". The final result is an unbalanced dataset to train and test your ML model. For the Santiago case the used grid dimension is 100x100m.

2. Grid and streets raw data structure. The grid is a discrete tool for splitting Santiago's surface but you must take in consideration that car crashes happened basically only on roads and streets, so the grid should consider only cells that contain streets. The join between primary grid and roads result in a subset of useful cells and a better balance between the amount of cells with crash zero (False) and one (True). The grid construction is a chain process: divide geography with 100x100 mt grid cells, joint the primary grid with OSM streets finally delete cells without roads.


2018 Real Crashes (Label) by Sensus Zone - Santiago Municipality
40
35
30
25
20
15
10
5
0


2018 Crash Predictions by Sensus Zone - Santiago Municipality
50
40
30
20
10
0

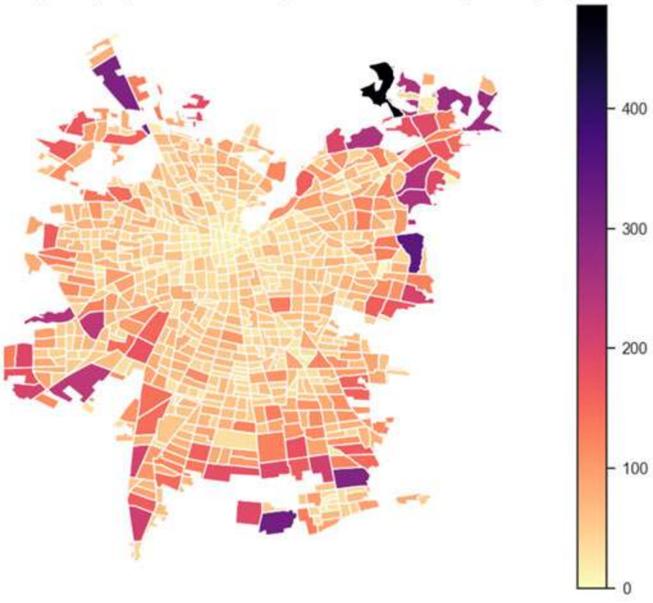
Car crash prediction advice



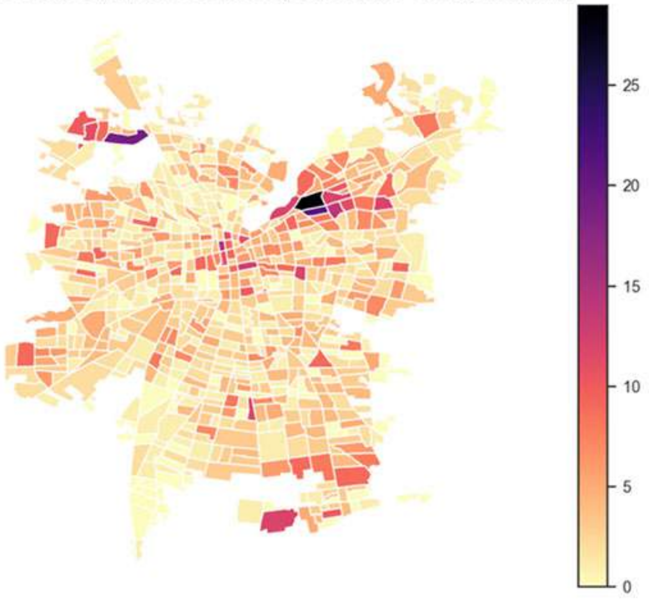
After the feature engineering process, the final dataset contained xx variables: 'X', 'Y', 'bank', 'bench', 'beverages', 'bus_stop', 'bus_stop_100', 'cafe', 'convenience', 'convenience_100', 'convenience_200', 'crossing', 'crossing_100', 'fast_food', 'fast_food_100', 'fast_food_200', 'fuel', 'intercect', 'kindergarten', 'motorway_junction', 'parking', 'parking_bicycle', 'pharmacy', 'railway_station', 'railway_station_100', 'restaurant', 'restaurant_100', 'school', 'school_100', 'school_200', 'stop', 'stop_100', 'taxi', 'traffic_signals', 'traffic_signals_100', 'turning_circle', 'ATROPELLO_100', 'ATROPELLO_200', 'CAIDA_100', 'CAIDA_200', 'CHOQUE_100', 'CHOQUE_200', 'COLISION_100', 'COLISION_200', 'INCENDIO_100', 'INCENDIO_200', 'OTRO TIPO_100', 'OTRO TIPO_200', 'SEV_Index_100', 'SEV_Index_200', 'VOLCADURA_100', 'VOLCADURA_200'.

Appart from staic and dynamic features, final dataset was enrich with cetroid of grid cell, and an atraccion area from each crash. This desition was a test feature seeking gain some balance to the final train/test ML dataset and because the hypotesis that crashes are not only explained by chance, but by bad infraestructure too.

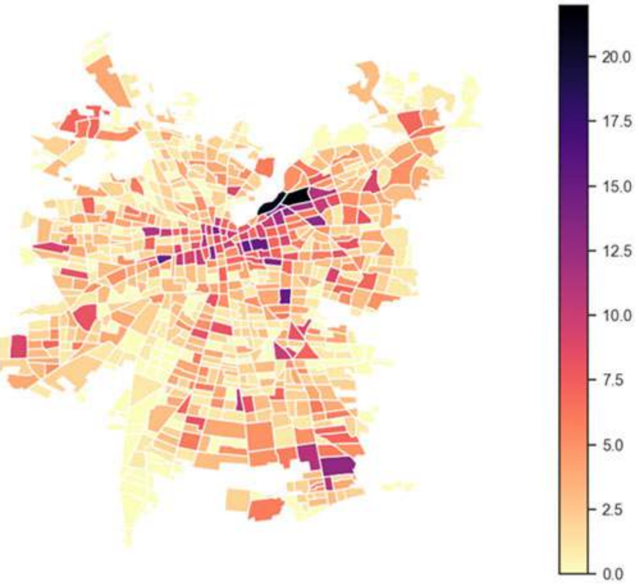
2018 True Negatives (TN) Crash Predictions by Sensus Zone - Santiago Municipality



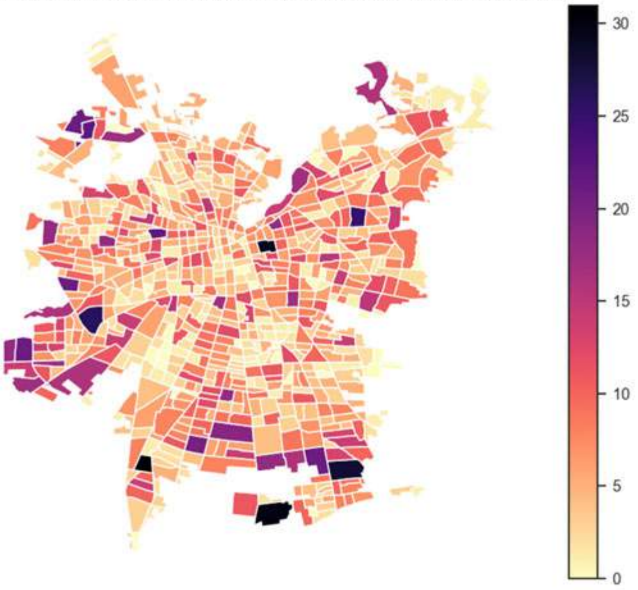
2018 False Positives (FP) Crash Predictions by Sensus Zone - Santiago Municipality



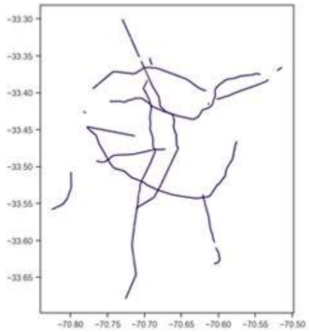
2018 True Positives (TP) Crash Predictions by Sensus Zone - Santiago Municipality



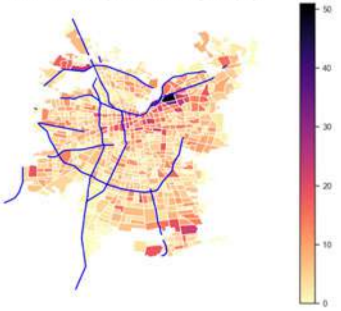
2018 False Negatives (FN) Crash Predictions by Sensus Zone - Santiago Municipality



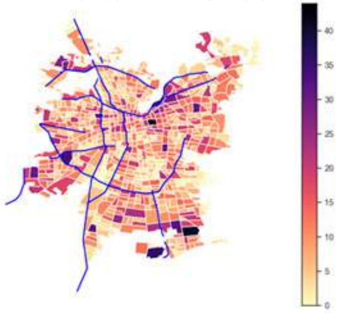
```
Autopista Central                       321
Autopista Vespucio Sur                  115
Autopista Costanera Norte               115
Autopista Vespucio Norte Express         97
Autopista del Aconcagua                  44
Avenida Presidente Kennedy               42
Autopista del Pacífico                   32
Autopista del Sol                        26
Autopista Acceso Sur                     14
Acceso Túnel San Cristóbal                5
Autopista Central - Eje General Velásquez 4
Autopista Acceso Sur, Tunel Sur           4
Autopista Acceso Sur, Tunel Norte         4
Autopista Aeropuerto                      3
Túnel Kennedy                             3
Salida Túnel San Cristóbal                2
Autopista Los Libertadores                2
Name: name, dtype: int64
```



2018 Crash Predictions by Sensus Zone - Santiago Municipality



2018 Real Crashes (Label) by Sensus Zone - Santiago Municipality



COMPARISON OF ML ALGORITHMS AND THEIR LIMITS

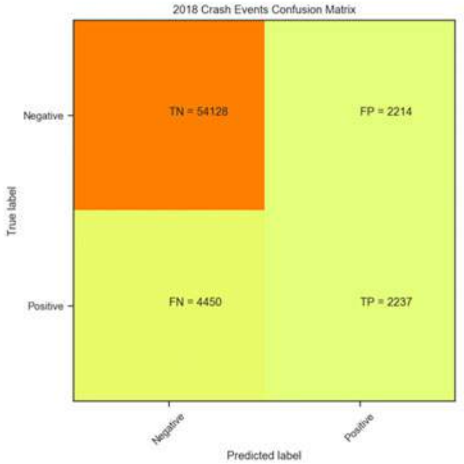
The ML Models: Random Forest (RM), GBT (Gradient Boosting Trees) and a scalable end to-end tree boosting system (XGBoost) were compared with three KPIs: simple Test Error (1.0 – accuracy), Test area under ROC (Receiver Operating Characteristic), and accuracy. The three models were fited with train/test set from 2013-2017 and evaluated their results:

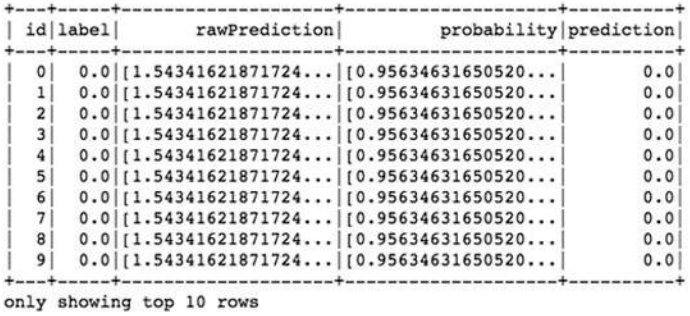
Random forest Classifier_ Test Area Under ROC= 0.8294667464568877
Random forest Classifier_ Accuracy= 0.7782577959311916
Random forest Classifier_ Test Error = 0.221742

Gradient Boosting Tree_ Test Area Under ROC= 0.8383690618564904
Gradient Boosting Tree_ Accuracy= 0.7879978006440971
Gradient Boosting Tree_ Test Error = 0.212002

Gradient Boosting Tree (Grid search)_ Test Area Under ROC= 0.841362042678081
Gradient Boosting Tree (Grid search)_ Accuracy= 0.7924750608750295
Gradient Boosting Tree (Grid search)_ Test Error = 0.207525

XGBoost_ Test Error = 0.210117 (Worst than the GBT with grid search)





```
+---+-----+--------------------+--------------------+----------+
| id|label|       rawPrediction|         probability|prediction|
+---+-----+--------------------+--------------------+----------+
|  0|  0.0|[1.54341621871724...|[0.95634631650520...|       0.0|
|  1|  0.0|[1.54341621871724...|[0.95634631650520...|       0.0|
|  2|  0.0|[1.54341621871724...|[0.95634631650520...|       0.0|
|  3|  0.0|[1.54341621871724...|[0.95634631650520...|       0.0|
|  4|  0.0|[1.54341621871724...|[0.95634631650520...|       0.0|
|  5|  0.0|[1.54341621871724...|[0.95634631650520...|       0.0|
|  6|  0.0|[1.54341621871724...|[0.95634631650520...|       0.0|
|  7|  0.0|[1.54341621871724...|[0.95634631650520...|       0.0|
|  8|  0.0|[1.54341621871724...|[0.95634631650520...|       0.0|
|  9|  0.0|[1.54341621871724...|[0.95634631650520...|       0.0|
+---+-----+--------------------+--------------------+----------+
only showing top 10 rows
```

For the final validation over the dataset 2018* with the GBT(Grid search). The results were:

Gradient Boosting Tree (2018 final test)_ Test Area Under ROC: 0.7742939707280319
Gradient Boosting Tree (2018 final test)_ Accuracy: 0.8942708911770773
Gradient Boosting Tree (2018 final test)_ Test Error = 0.105729

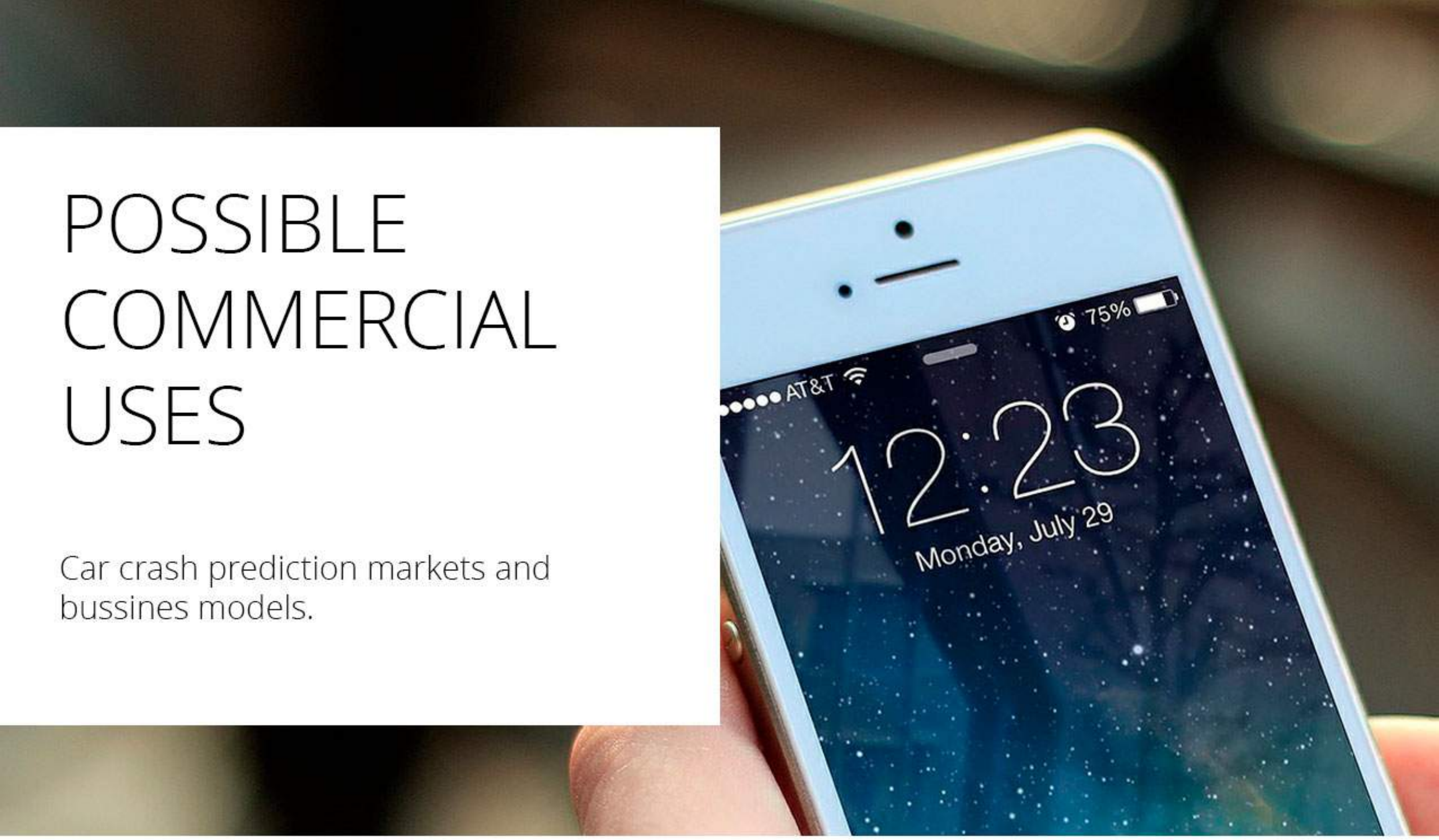
These results were obtained tuning the hyper parameters as follow descripted:

- number_of_trees 58
- number_of_internal_trees 58
- model_size_in_bytes 45603
- min_depth 6
- max_depth 6
- mean_depth 6.0
- min_leaves 42
- max_leaves 64
- mean_leaves 57.9310

* Crashes from 2018
Dynamic features from 2017

POSSIBLE COMMERCIAL USES

Car crash prediction markets and bussines models.



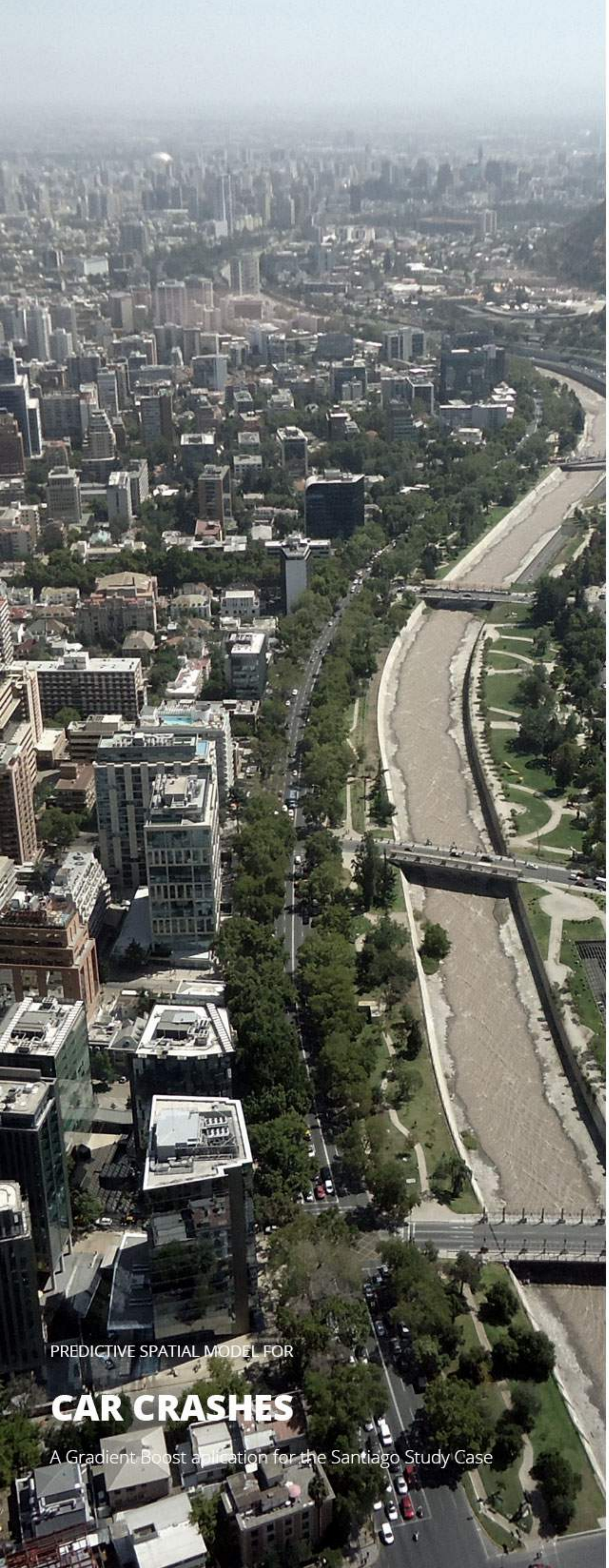
Even it’s possible to recommend a less risky route

There are several commercial uses for the space car crash predictive model. Car kilometers insurance are new business models with behavioral driver features but without urban risk route scoring. The model can built an aggregate risk scoring for driver’s routes. Even it’s possible to recommend a less risky route based on aggregate possible trips comparison.

Following this kind of business models, B2B services as logistic transport or final mail/retail delivery might use a risk scoring model to prevent over cost do truck crashes or other kinds of accidents that might result in high costs or sues.

Public or private massive transport. Most of the real time routes built by apps only consider a cost versus time measure. This is might be a political issue for public transport or maybe for on demand car sharing and car-pooling as security and safety strategy for massive transport.

CONACET dataset is not only for car crash. There are data of pedestrian abuses with different results of injury. The model would be trained as well with this data opening commercial uses for people life insurance too.



RESULTS AND FUTURE WORKS

The final user must have a real time engine.

All ML models tested are good performers in the experiment. GBT with grid search wins over the three KPIs measured. Despite this all the models are closed in their results. Finally the case had an 89% accuracy with problems from unbalance dataset and no time and Origin-destiny features.

CONACET dataset is not only for car crash. There are data of pedestrian abuses with different results of injury. The model would be trained as well with this data opening commercial uses for people life insurance too.

The final enrich data set (Static and dynamic joined data) must be re- enrich by 'origin_destiny' social demographic data (Possibly TELCO data). Cells that are destiny of trips or origin of mobility, cells that are highly dense or poorly dense might have a great impact in the feature importance list. some extra features:
+Average road speed
+Demographic density
+Mobility factor
+Roads geometry (curvature, inclination, etc.)
+Altitud
+Number of road lanes
+Proximity with highways/main roads
+Other

Time of event could be a powerful feature for prediction. Unfortunately CONACET data set has not date time link to any crash. A new source must be found.

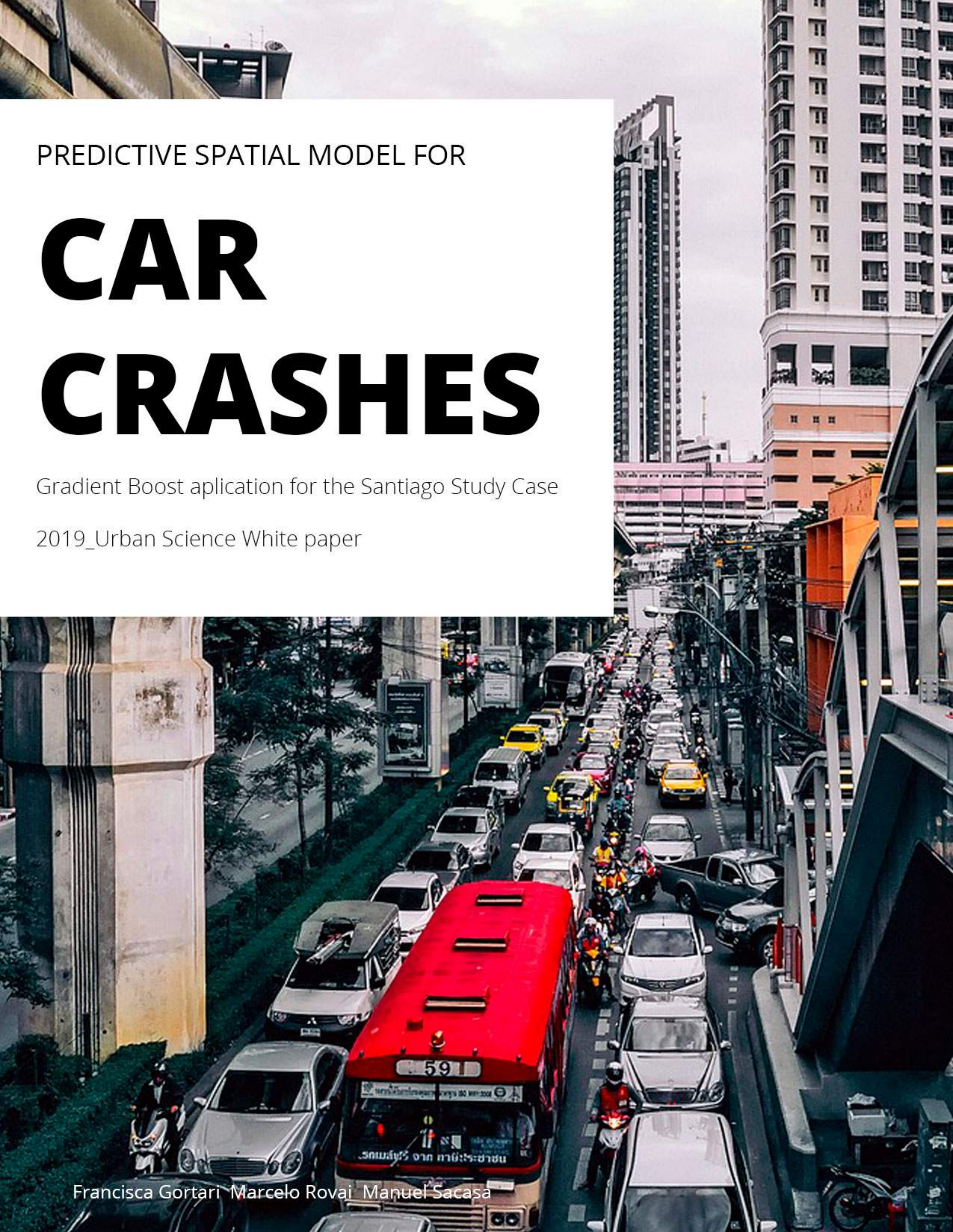
The final user must have a real time engine in a channel like an app, bot or other interface to query in real time the risk score in a user centric designed product/service.

PREDICTIVE SPATIAL MODEL FOR

CAR CRASHES

Gradient Boost aplication for the Santiago Study Case

2019_Urban Science White paper



Francisca Gortari Marcelo Rovai Manuel Sacasa