Read_Me

Car crash prediction

"Modelo predictivo (Espacial) de siniestros en las calles de Santiago"

UDD - Universidad del Desarrollo
MDS-18
BDA
Peredo, Oscar
Professor

2019-08-09

Gortari, Francisca
(https://www.linkedin.com/in/franciscagortari/)
Rovai, Marcelo
(https://www.linkedin.com/in/marcelo-rovai-2186762/)
Sacasa, Manuel
(https://www.linkedin.com/in/manuel-antonio-sacasa-ares-63b66b77/)
Master Candidates

The car crash prediction research for Santiago de Chile is a final exam work for the Big Data Analytics course of the UDD's (Universidad del Desarrollo) Data Science Master Degree.

The final objective of the research is to predict a crash risk score for an urban grid with a 2013-2018 car crash georeferenced dataset and static urban descriptive public data set. The model requirement was to train and test a model using Extreme Gradient Boost algorithm and compare the results against another ML algorithms.

A_Files

In the folders structure, you will find:

01_Read_me: brief description of the research, data, pipeline and results (PDF file).

02_CCP_Car Crash Prediction: powerpoint slides of the final presentation describing the details of the research, data, pipeline and results (PDF/PPT file).

03_White_Paper: Urban science wrap up, georeference grid tips, crash prediction advice, limits of the XGB, warnings, possible commercial used and future research (PDF file).

04_Final_ Code folder:

data (Folder)
- CONASET (Folder)
- OSM_Chile (Folder)
- Final_test_dataset_grid_100.csv
- Final_train_dataset_grid_100.csv
- Geo_stgo_test_crash_test_100.csv
- Geo_stgo_test_crash_train_100.csv
- Geo_stgo_100_estatic_dataset.csv

model (Folder)
- GeoProjectBestModel_1.model (Folder)
- Crash_Prevision_2018.csv
- Visualization_Crash_Prevision_2018.csv

notebooks (Folder)
- 10_Final_Geo_Project_Grid-creation_Static_feature. Geographical division of territory in a geometric grid and enrichment with static descriptive features of the environment.
- 20_Final_Geo_Project_Dynamic_Features
- 30_Final_Geo_Project_Finaldataset_Creation
- 40_Final_Geo_Project_PySpark_MLib_RF_GBT_Models_Gr100_dataset-2013-17_val_18
- 41_Final_Geo_Project_PySpark_MLib_XGBoost_Model_Gr100_dataset-2013-17
- 50_Final_Geo_Project_Model_Validation_Analysis
- 60_Final_Geo_Project_EDA_Model_Result_Visualization

05_Image (Folder). Images from map plots of raw dataset and results.

B_Data

1_CONASET (http://mapas-conaset.opendata.arcgis.com/search?groupIds=fca1f61c6556499db843c09cc80c70c0)
6 Datasets. Georeferenced car crashes from 2013 until 2018 enrich with injury, type of crash, wounded persons, etc. Name of the datasets "Siniestros RM20XX".

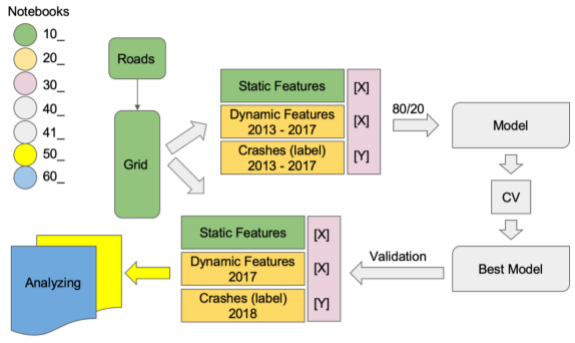
2_OPEN STREET MAP _ OSM_Chile
(http://download.geofabrik.de/south-america/chile.html)
"Chile-Latest-free" data set with georeferenced points: places, points of interest, points of worship, natural, traffic, transport. With line strings: roads, railways and waterways. With polygons: buildings, land use and water. View "OpenStreetMap Data in Layered GIS Format" file. EOD_2012 is an extra data set in OSM folder containing the mobility survey of Santiago (This survey is the official dataset built with the accepted methodology).

3_CENSO Chile 2017.
(http://www.censo2017.cl/microdatos/)
Georeferenced dataset of the inhabitant social survey from Chile (2017). This data set's features are total population, Social economic clusters and other social descriptive variables. Has no mobility features. The smallest geographical unit is a block and all the samples are aggregated and anonymized. (To be used on a following work).

C_Notebooks Pipeline, tools and libraries

All the research was processed in MacBook Pro (Retina, 15-inch, 2017) with 2.9 GHz Intel Core i7 and 16 GB 2133 MHz LPDDR3 Memory, processing power on-premise machines.

The data science coding environment was Anaconda's Jupyter notebooks running Python 3.7.1 and PySpark 2.4.3. Maps, and Grids were developed mainly with GeoPandas 0.4.1, supported by several packages as Folium and OSMNX.



The end to end "notebook's pipeline" is:

10_Final_Geo_Project_Grid-creation_Static_feature.

- Geographical division of territory in a geometric grid (Using Santiago municipality shape file), jointed to streets (CD - "Calles Discretizadas"). A generic function was developed and grids of 50mx50m and 100mx100m were generated.
- Development of functions to feature creations.
- Dataset creation of static descriptive features, based on georeferenced points: places, points of interest, natural, traffic, transport., intersections and roads. (OSM_Chile).

20_Final_Geo_Project_Dynamic_Features.

- Development of functions to create and aggregate dynamic features to dataset, based on historical car crash influence (based on distance from grid cells and year of occurence.
- Dataset creation with dynamic features (data from CONASET dataset)

30_Final_Geo_Project_Final_dataset_Creation.

- Dataset creation Jointing data from dynamic and static features.
- Adding Grid centroids as features
- Splitting Dataset on a "Train dataset", to be used on ML model development. This dataset will use dynamic features generated crashes from 2013 to 2017. Dependent variable will be related to those events.
- Splitting Dataset on a "Test dataset", to be used as ML model's validation. This dataset will use dynamic features generated crashes from 2017 only. Dependent variable will be related by 2018 crash events.

40_Final_Geo_Project_PySpark_MLib_RF_GBT_Models_Gr100_dataset-2013-17_val_18. T

- Development of models, based on RF (baseline) and Gradient Boosted Tree (GBT) algorithms with the train dataset (2013-2017).
- Tuning GBT Model (CV)
- Validating Best Model with Test dataset (2018)
- Saving best Model

41_Final_Geo_Project_PySpark_MLib_XGBoost_Model_Gr100_dataset-2013-17.

- Training and testing an Extreme Gradient Boosted (XGBoost) algorithm with the train dataset (2013-2017).

50_Final_Geo_Project_Model_Validation_Analysis

- Analysing the Crash predictions generated by best model.
- Confusion Matrix and error report generation

60_Final_Geo_Project_EDA_Model_Result_Visualization

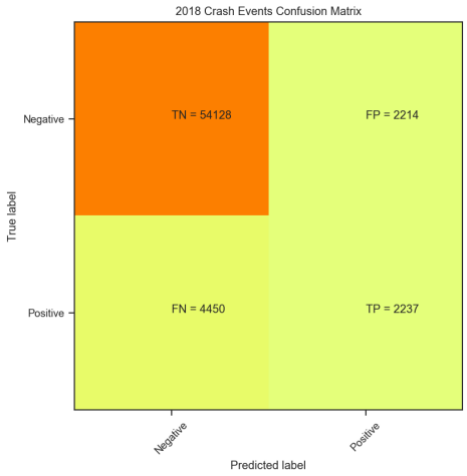
- Spatial visualization of raw dataset (from 2013 to 2018) and results from model predictions.
- Geo Visualization of Confusion Matrix using as context Santiago Municipality and Sensus zones.

The next list are the libraries used in the end to end pipeline jupyter notebooks:

- Math
- Numpy
- Pandas
- Pylab
- Seaborn
- Folium
- Networkx
- Osmnx
- Geopandas
- Shapely
- MultiLineString
- Shapefile
- Gpd_lite_toolbox
- Findspark
- Pyspark

D_Results

	precision	recall	f1-score	support
0.0	0.92	0.96	0.94	56,342
1.0	0.50	0.33	0.40	6,687
accuracy			0.89	63,029
macro avg	0.71	0.65	0.67	63,029
weighted avg	0.88	0.89	0.88	63,029



E_Bibliography

- CONASET- Análisis espacio temporal de los siniestros de tránsito en el Gran Santiago. Diagnóstico 2009 -2013
- Traffic Accident Analysis Using Machine Learning Paradigms, 2005, Miao Chong et al.
- Using Machine Learning to Predict Car Accident Risk, Daniel Wilson May 3, 2018
- Big Data Analytics, UDD/MDS18 2019, Oscar Peredo
- PySpark ML and XGBoost full integration tested on the Kaggle Titanic dataset
- Machine Learning with PySpark and MLlib — Solving a Binary Classification Problem
- Build an end-to-end Machine Learning Model with MLlib in pySpark.
- GeoPandas main functions Based on the work of Eduardo Graells-Garrido