Machine Learning Engineer Nanodegree

Capstone Proposal

By Jorge Bartra
June 27th, 2018

Domain Background

The project is a real world scenario for a Non-Profit organization. The organization is facing a challenge today because there are more than 100K potential donors that are currently not being included by the Marketing department in their campaigns. The 100K potential donors has been overlooked or not used because there is a lack of knowledge in regards to Machine Learning skills.

Currently the Non-Profit organization requires additional source of income in order to generate extra revenue and expand charity missions around the globe and also support and increase disaster recovery campaigns.

Machine Learning is been used in all types of business. Non-profit organizations are not stepping away from Machine Learning. Rather they are using it to increase the amount of donors by executing marketing campaigns constructed on Machine Learning algorithms. These campaigns are based on certain demographics such as particular interest, financial status and some others (1).

This is the reason why this project will not be the first one on its class since Machine Learning has been used already to increase the donation to Non-Profit organizations.

The project is a real scenario and I have been approved by Udacity team to continue with this task

Problem Statement

There are different sources of data distributed in different servers and databases in the organization. Currently this data is in silos and is not efficiently connected. Using a series of SQL Stored Procedures the data can be combined into one single temporary table and stored for further analysis. Using Stored Procedures will save time next year since is highly probable that the model will be executed again.

The data presents a challenge because the data does not have quantitative features that could be used in the model such as “Income”, therefore it is necessary to obtain data from a public domain that after is joined to the organization’s data, it could produce a quantitative feature for every single record that is obtained within the organization’s databases.

The model planned to be use is a Regression Model since the defendant variable is continuous. The output desired for this model is the Mean Income/ ZCTAs. After optimization the desired Accuracy and F-score would be close to 85% (+-).

Datasets and Inputs

There are 4 sets of data located in different servers and databases and all of them will be used in this project. Two of the servers are SQL servers located in different locations within the US. The third and fourth datasets are located on a public domain.

The first dataset is composed of two different SQL tables. The first table contains the CONSTITUENT SYSTEM ID (Unique Key) and it will be used in the model rather than personal information such as Name and Last Name. The second table has the CONSTINUENT SYSTEM ID as a foreign key and it has the filter needed to exclude the rest of the CONSTITUENTS who either already donated funds or are already future prospective donors.

The dataset collected has several features that can used in the model such as gender, age and address. Within those features there are 4 (Address, City, State, and Zip Code) fields that can be used as a composite key field to join to the external data and obtain the quantitative field needed by the model. The public domain data are stored in two different websites.

Census Data

https://www.census.gov/programs-surveys/decennial-census/data/datasets.2010.html

Redfin Data Center

https://www.redfin.com/blog/data-center

Data massaging and preparation is needed before it can be used in the model. After the public domain data is ready, it needs to be downloaded into a local drive in a readable format such as CSV or TXT extension files.

After the data is located in the local drive an ETL process (SSIS) will be constructed to join public domain data with the data warehouse. After that the model can be utilized.

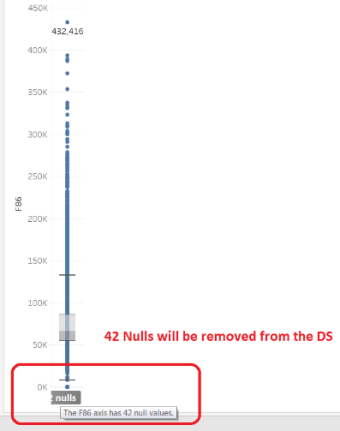
The number of rows in each dataset is as follows:

- Combined table using SQL servers: 102510
- Census Data: 14519 rows (These rows are aggregated at the ZCTA level)
- RedFin Data Center: 15700 ((These rows are aggregated at the City and Zip Code level)

There are 14519 distinct target variables but when this dataset is joined to the SQL table it will be 102510 target variables since the SQL Server data is at the lowest level (SYSTEM ID or potential donor level)

For categorical variables, one-hot encoding algorithm seems to work very well for the type of features the model will use.

The dataset distribution seems to be balanced and it will be analyzed using a simple Box Plot graph or using python scripts. In the screen shot below they is a simple Box Plot. The first thing to do is remove the nulls:



After removing the NULLS these are the results of the target variable:

Upper Whisker:	132,466
Upper Hinge:	85,594.5
Median:	66,502.5
Lower Hinge:	54,336
Lower Whisker:	8,223

Additional analysis is required to find if the data is skewed or not and then Logarithmic transformation will be applied.

Solution Statement

Since the organization's data does not have quantitative fields, the solution is to join the organization's data with the public domain's data so the model could use a quantitative feature.

The public data has the house mean sale price and mean income by ZCTAs or zip codes. Incorporating the quantitative features into the data will allow the model to predict the potential donors with higher probability and be targets for the marketing campaigns.

The organization's data has the house addresses with city, state and zip code. These three fields will be used as composite key and it will be used to join the public domain data. If the join results are too complex, only the zip code will be used.

Zone Improvement Plan or ZIP codes would not be the preferred field to use when the data is prepared. ZIP codes are based on group of addresses that make the mail delivery efficient but the model will use CENSUS data. Therefore instead of using ZIP codes, the model will use ZCTAs which in some cases are the same as ZIP codes but in some other cases are represented for larger boundaries on the maps (2).

In conclusion: there will be a house mean sale price and mean income by ZCTAs (5 digits code). The mean sale price and mean income will be used as the quantitative features in the data.

Benchmark Model

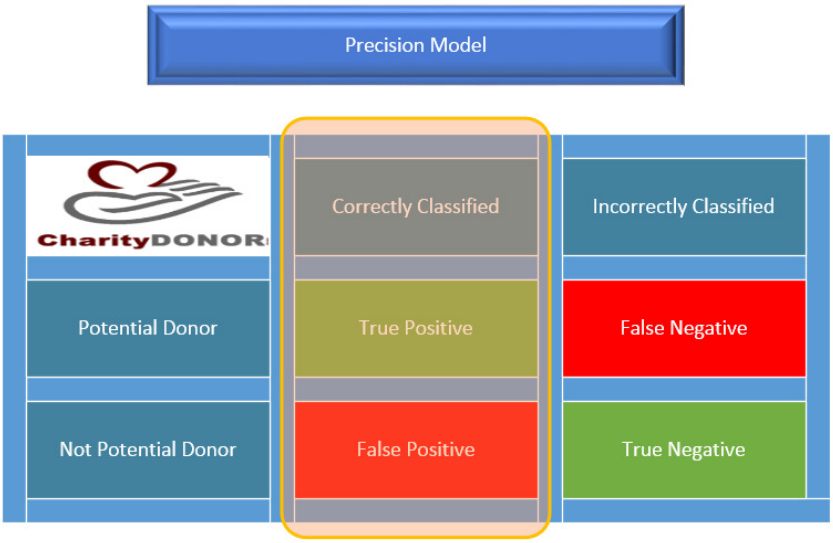
Since the data is not at the individual level but rather at the zip code and city level, the organization is interested in targeting individuals where the house mean sale price is higher than 180K.

In addition to the house sale price the estimated mean income by zip code will be also used. The targeted individual who make an average higher than 70K/year.

Using the above target the model to use will be Simple Regression Model

Evaluation Metrics

The intention is to use Precision Model rather than Recall. Below is the confusion matrix drawing elaborated for this project:



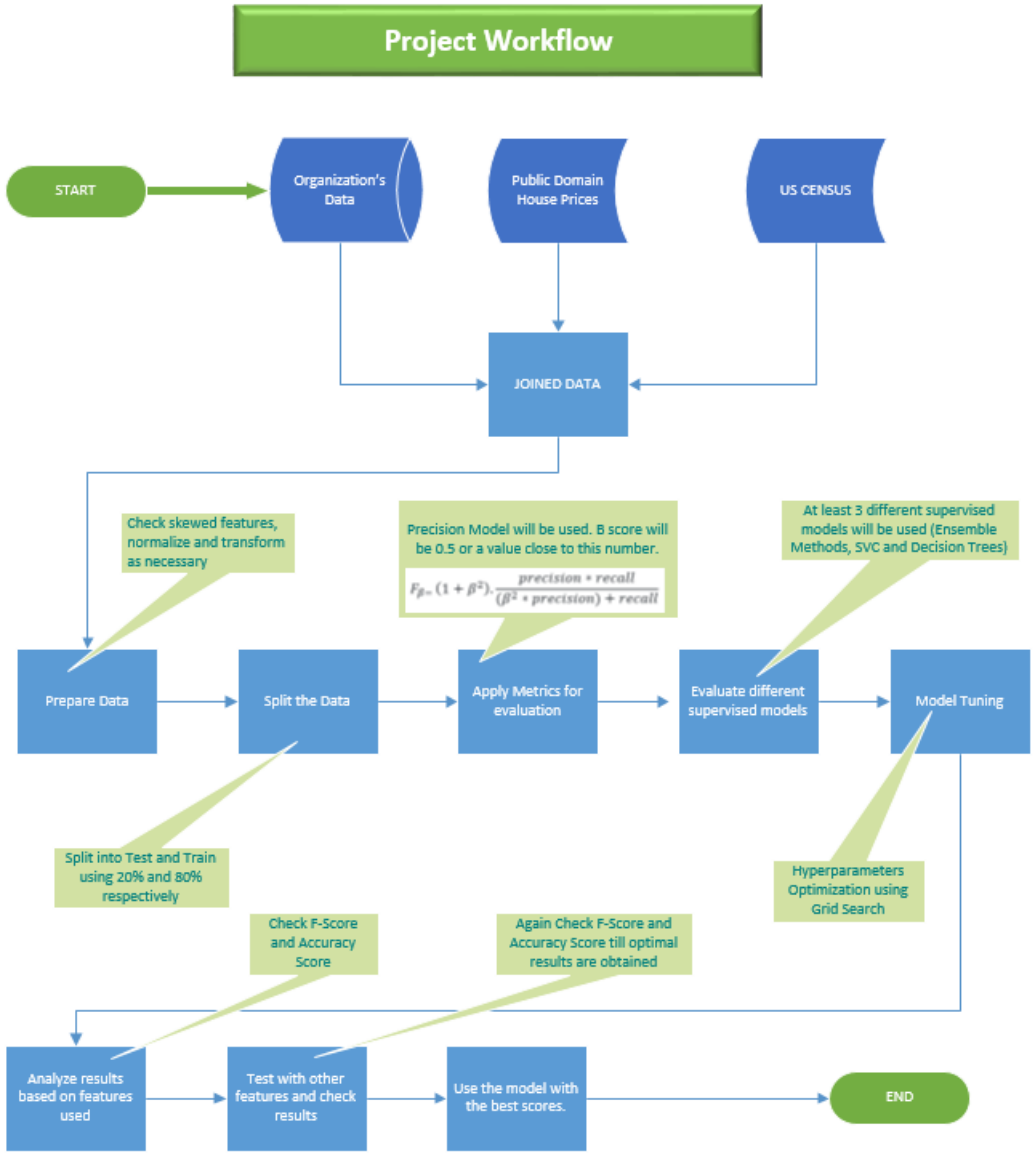
$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Since it is a Precision model then it is necessary to use the B score with a value of 0.5 using the below formula:

$$F_{\beta=}(1+\beta^2).\frac{precision * recall}{(\beta^2 * precision) + recall}$$

The model will use other features such as education, age, gender, marital status, household income, etc. therefore the intention is not to drop everyone that makes higher than 75K/year but rather combined all the most relevant features in the data and obtain the best F score and Accuracy values.

Project Design



Reference

1. http://blog.abila.com/artificial-intelligence-machine-learning-nonprofits-associations/
2. https://www.census.gov/geo/reference/zctas.html