Machine Learning Engineer Nanodegree

Capstone Proposal

Brindha Sivashanmugam

August 6th 2018

EMERGENCY – 911 CALLS PREDICTION

(Dataset: Emergency – 911 calls in Montgomery, Pennsylvania)

DOMAIN BACKGROUND

We all know how significant the role of a 911 employee is. They are the ones who address our emergencies. The team works 24x7 non-stop. Any challenges they face in their job will directly affect the public in their life or death moment. Let us look at what happens when there is a flaw on their side, how it affects the public, and the challenges they face in their day-to-day job.

From the Headlines: (How a flaw in 911 affects the Public)

Longer Response Time: Tragic death of Shanell Anderson in Atlanta because emergency responders arrived too late to save Shanell Anderson from drowning in her SUV. *(Reference #1)*

Equipment Outage: A cooling unit low on refrigerant, caused Montgomery County, MD 's 911 system to crash for nearly two hours on the night of July 10 2016, causing about 100 callers to get busy signals when they tried to report emergencies. Delays caused deaths of a 40-year-old dialysis patient and a 91-year-old woman. *(Reference #2)*

Human error: The death of Kokomo resident Tammy Ford on July 1 2015, due to operator error. In that case, a dispatcher sent first responders to the wrong address. *(Reference #3)*

Challenges Faced by 911 Operators and Dispatchers:

- Most of the 911 employees work in 12 hour shifts in a hectic, high stress environment, not able to plan their leisure hours or vacation. They are humans too. If they could know how to plan this, they could be stress free, and operator error will no longer be an issue.
- Not able to know when they could schedule a downtime to maintain their equipment, which is very crucial, so that the equipment would to be in good shape, ready for a dire emergency.
- Not able to predict the prime location where most of the calls are expected on a particular day and time. If they had known that, they could rotate employees accordingly and cover the emergencies in a shorter response time.

Research in such area has already been started, and the academic paper "Mining 911 Calls in New York City: Temporal Patterns, Detection and Forecasting" is an example. *(Reference #4)*

PROBLEM STATEMENT

If the response time of the 911 team is minimized, lot of lives could be saved. Response time can be minimized if there are surplus amount of equipment and human resources ready to serve emergency, which is practically not possible, as it would consume a lot of money.

So, the alternate way is to effectively distribute available resources. This could be possible only when they know in advance about how many 911 calls are expected. As this is emergency call, no one knows what is expected until it actually happens. Here is where Machine Learning comes in rescue of the rescue team.

The problem would be a regression. Instead of predicting the number of emergencies expected overall, I have planned to predict 3 variables - the number of emergencies expected in each category - EMS, Fire and Traffic. The inputs would be the geographic location, date and time. The dataset does not explicitly include a column to indicate the count per emergency category. So, it will be rearranged accordingly.

DATASETS AND INPUTS

I am using "Emergency - 911 calls" dataset for this problem. This dataset was provided by montcoalert.org, and donated to Kaggle by Mike Chirico. This dataset contains more than 2 years history of 911 calls handled by Montgomery county, Pennsylvania. It contains the emergency category, location and GPS coordinates of the victim. As it is very informative, and the dataset has enough values to do the analysis, I thought I could work on this for my Machine Learning Capstone Project. I have downloaded the dataset from Kaggle Datasets. *(Reference #5)*.

The dataset has 326425 records and 9 columns. The columns are described below:

S.No	Column Name	Column Description
1.	"lat"	Latitude
2.	"lng"	Longitude
3.	"desc"	Description of emergency
4.	"zip"	ZIP code
5.	"title"	Title of emergency
6.	"timeStamp"	Date and time of the call
7.	"twp"	Town
8.	"addr"	Address
9.	"e"	This column is filled with a constant 1 for all rows

The below is a sample row from the dataset.

```
data.head(1)
```

	lat	lng	desc	zip	title	timeStamp	twp	addr	e
0	40.297876	-75.581294	REINDEER CT & DEAD END; NEW HANOVER; Station ...	19525.0	EMS: BACK PAINS/INJURY	12/10/2015 17:10	NEW HANOVER	REINDEER CT & DEAD END	1

SOLUTION STATEMENT

With the available timestamp and location information, a regression model would be developed which predicts the number of calls in each emergency category on a given day, time window and location. Time window will be explained in detail in the “Project Design” section.

XGBoost model*(xgboost.XGBRegressor)* will be implemented for predicting the number of calls. Evaluation is done based on R2 score, which will be explained under “Evaluation Metrics” section.

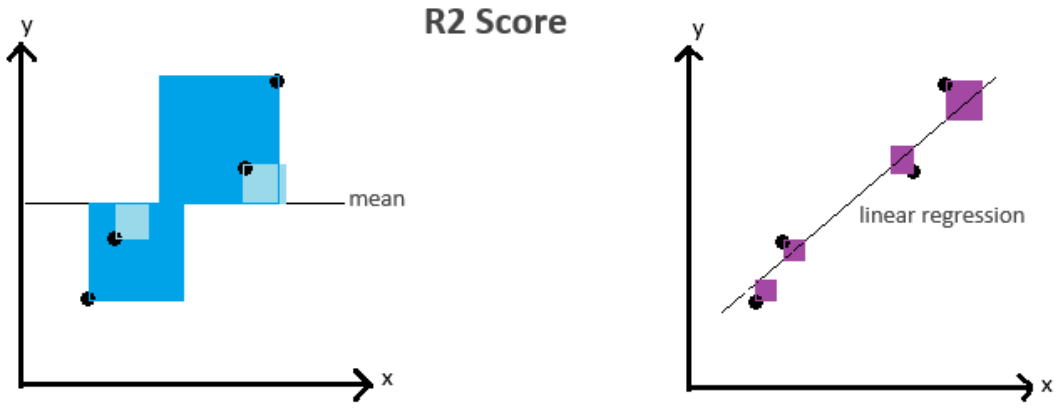
BENCHMARK MODEL

Linear regression model *(sklearn.linear_model.LinearRegression)* will be implemented to predict the number of 911 calls expected in each emergency category on a given day, time window and location. This model will be evaluated using R2 score as the evaluation metric. This model will be used as a benchmark model.

EVALUATION METRICS

R2 score is used for evaluating all the models developed in this project. R2 score is a very common regression metric. It is based on comparing our model to the simplest possible model (the mean of the data points).

This will be clearly given in the below plot.



The areas of the blue square represent the squared residuals with respect to the mean.

The areas of the purple square represent the squared residuals with respect to the linear regression.

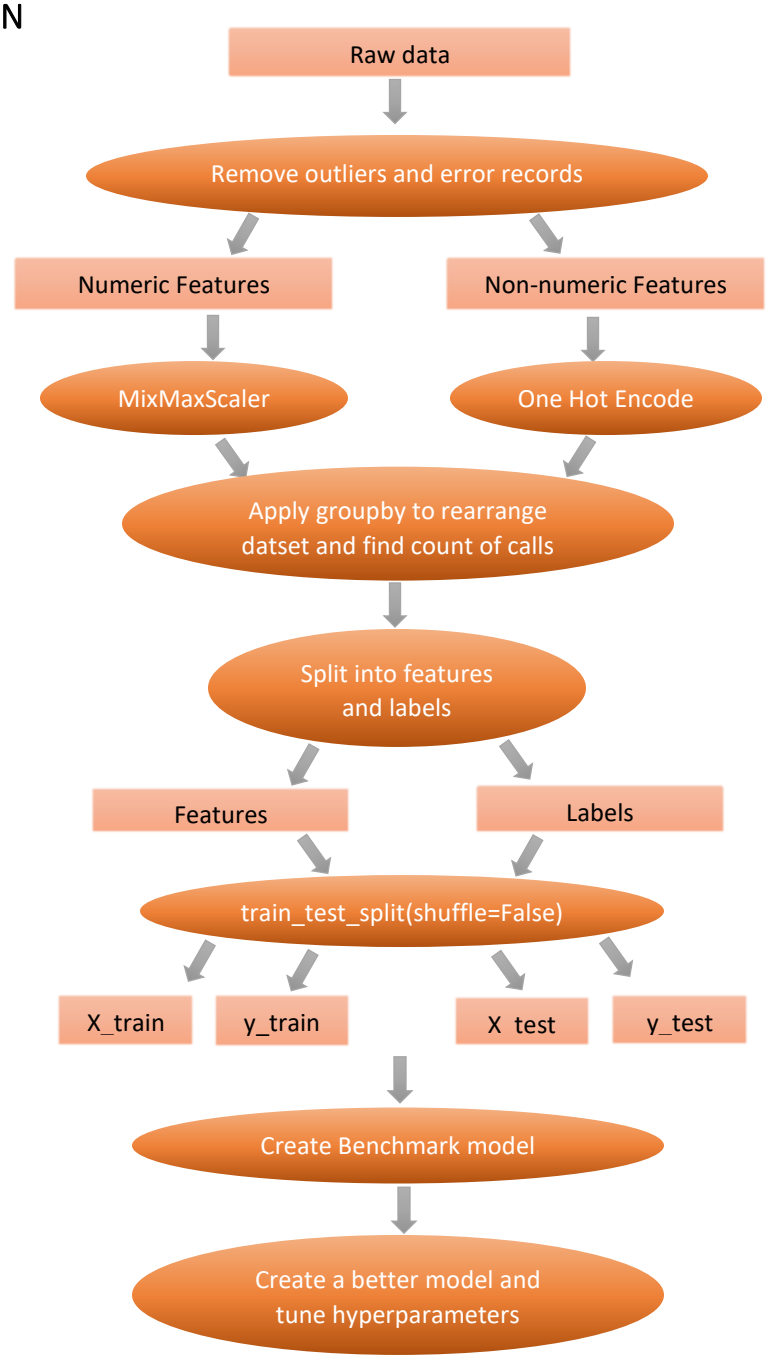
$$R2\ score = 1 - (SS_{linear\_reg} / SS_{mean})$$

The simplest possible model is the average of all data points, represented as a horizontal line that goes through them. The sum of squares of residuals of this simple model is calculated (SS_{mean}). And, the sum of squares of residuals of our model, for example, linear regression model is calculated (SS_{linear_reg}). Then, R2 score would be defined as:

$$R2\ score = 1 - (SS_{linear\_reg} / SS_{mean})$$

The value of R2 score lies between 0 and 1. In order for a model to be good, the R2 score should be closer to 1.

PROJECT DESIGN



Data cleanup

1. “twp” column will be checked if all towns are Montgomery county’s towns. If any other values are identified, they are outliers and will be removed.
2. Latitude and longitude coordinates of the Montgomery county edges will be obtained from Google Maps. If any data has a latitude – longitude value far out of range, they must be error records which happened due to error in operator data entry. These error records will be identified and removed.
3. Columns contain missing values will be identified and removed. Also, columns containing redundant information will be identified and removed.

Data Preprocessing

1. Using the latitude, longitude bounds of Montgomery county, grid will be formed to separate it into regions, and this region will be added as a feature.
2. Timestamp column will be split into day, week, day of week, month, year and time.
3. Time will be replaced by time window. 3 time_windows would be considered (values: 0, 1, 2) which means:
    a. 0 => 12AM to 8AM;
    b. 1 => 8AM to 4PM;
    c. 2 => 4PM to 12AM
4. Data needs to be scaled before applying learning algorithms on them, so that each feature will be treated equally by the learning algorithm. So, MinMaxScaler will be applied to the numeric features.
5. One Hot Encode categorical features (region & emergency title)

Re-arranging the dataset

Though there are a lot of emergency types under “title” column, they broadly fall under 3 categories – EMS, Fire and Traffic. We need to find the number of emergencies in these categories when all other information in remaining columns are the same. So groupby needs to be applied to rearrange the dataset to explicitly get the count.

Split the Dataset

The dataset will be split for features and labels. The number of calls per Emergency title (EMS, Fire, Traffic) will be the labels. Then, the dataset will be split in order (as this is time-series data) for training and testing. 70% of the data will be used for training, and 30% will be used for testing.

Setting the Benchmark

A linear regression model will be developed and applied, and the R2 score of the testing set will be recorded. This will be used as the benchmark score.

Developing a better model

xgboost regressor will be developed and the hyperparameters will be tuned to get a good model. R2 score of the testing set will be recorded.

REFERENCES

1: https://www.ravemobilesafety.com/blog/911-system-failures

2: http://www.kokomotribune.com/news/dispatchers-discuss-challenges-of-the-job/article_85753cb0-379a-11e5-8a06-b74ef6ca24dc.html

3: https://www.washingtonpost.com/local/md-politics/montgomery-officials-try-to-explain-911-outage-that-should-never-happen/2016/08/02/798c77d8-58d2-11e6-9767-f6c947fd0cb8_story.html?utm_term=.cb08532efcaf

4 https://warwick.ac.uk/fac/sci/dcs/people/theo_damoulas/pubs/aaai15_911damoulas.pdf

5: https://www.kaggle.com/mchirico/montcoalert