PredictingPedestrianVolumes

September 22, 2018

Machine Learning Engineer Nanodegree
Capstone Project
Tim Black
September 22, 2018

1 Definition

1.1 Project Overview

In order for cities to plan transportation systems, they need to have robust data on the number of people using different modes of transport. This is especially important when the urban landscape changes, such as an increase in urban density, and cities need to plan for the changes in traffic that will arrive. Typically, cities collect data on the volume of vehicle, bicycle, and pedestrian traffic. Unfortunately, collecting these data is costly and time consuming, and in some cases impossible. Often transportation engineers and planners need to understand the consequences of different design decisions, such as the estimated effect of a traffic signal on safety. Since it is impossible to build out these alternatives and then perform volume counts, engineers and planners need a way to be able to estimate the effect of changes on pedestrian volume.

Although many cities have models to estimate the effect of the built environment on pedestrian volume, they are typically at the scale of traffic analysis zones, a geographic scale much larger compared to the intersection. Intersection-based models exist for the following locations: San Francisco, CA (1,2); Charlotte, NC (3); Alameda County, CA (4); San Diego County, CA (5); Santa Monica, CA (6); and Quebec (7). Most of these intersection-based models use either a linear model, and the most common features found to significantly affect pedestrian volumes include population density, employment density, and transit accessibility.

Despite general agreement on the most important features, there are differences among the models on other significant features from the built environment. For example, the City of Santa Monica found the distance from the ocean to be a significant feature for prediction (6); it is highly unlikely that a landlocked city would find the same to be the case. Even when the models agree on which features of the built environment are significant predictors, they often disagree on the extent to which they influence pedestrian volume. As suggested by Schneider et al., this variation should be addressed by creating models that are sensitive to the context of the local environment (1). Since there currently does not exist a model to predict pedestrian intersection volumes for the City of Los Angeles, this project aims to fill that gap.

1.2 Problem Statement

For this project, I will answer the following questions:

1. Which features of the built environment are important in predicting pedestrian volume in Los Angeles?
2. What is the best modelling method to predict daily intersection pedestrian volume, and how well can this model predict pedestrian volume at an intersection?

1.3 Metrics

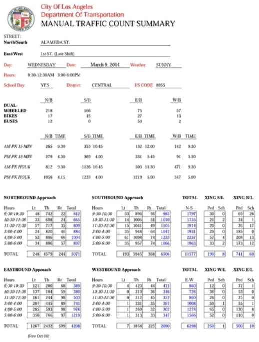
My proposed metric for evaluating my linear model is the r-squared score, the proportion of the variance in the dependent variable that is predictable given the features.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

2 Data Assembly

The Los Angeles Department of Transportation (LADOT) routinely collects volume data related to bicyclists, pedestrians, and motor vehicles for the purposes of transportation planning. Historically, these data have been stored in a PDF format, which makes it easy to digest a single traffic count, but prevents comparison and analysis of multiple counts. These PDF files are publicly available on the Navigate LA portal at http://navigatela.lacity.org/navigatela/. The first step to solving the proposed problem required extracting pedestrian volume from these sheets and storing it in a format that could be used in a model.

I developed the following python pipeline below to read the pedestrian volume data from the PDF sheets and format them so they can be used in building the model. The daily volume is represented by the 'volume' attribute in the table. Each row is a separate count event (`count_id`), which occurs at an individual intersection (`ASSETID`, `cl_node_id`).



City Of Los Angeles
Department Of Transportation
MANUAL TRAFFIC COUNT SUMMARY

STREET:
North/South ALAMEDA ST.

East/West 1st ST. (Late Shift)

Day: WEDNESDAY Date: March 9, 2014 Weather: SUNNY

Hours: 9:30-12:30AM 3:00-6:00PM

School Day: YES District: CENTRAL I/S CODE 8955

	N/B	S/B	E/B	W/B
DUAL-WHEELED	218	166	75	57
BIKES	17	15	27	13
BUSES	12	0	50	2

	N/B	TIME	S/B	TIME	E/B	TIME	W/B	TIME
AM PK 15 MIN	265	9.30	353	10.45	132	12.00	142	9.30
PM PK 15 MIN	279	4.30	369	4.00	331	5.45	91	5.30
AM PK HOUR	812	9.30	1126	10.45	503	11.30	471	9.30
PM PK HOUR	1058	4.15	1233	4.00	1219	5.00	347	5.00

NORTHBOUND Approach					SOUTHBOUND Approach					TOTAL	XING S/L		XING N/L	
Hours	Lt	Th	Rt	Total	Hours	Lt	Th	Rt	Total	N-S	Ped	Sch	Ped	Sch
9:30-10:30	48	742	22	812	9:30-10:30	33	896	56	985	1797	30	0	65	26
10:30-11:30	33	608	24	665	10:30-11:30	14	1005	51	1070	1735	21	2	34	1
11:30-12:30	57	717	35	809	11:30-12:30	15	1041	49	1105	1914	20	0	76	17
3:00-4:00	24	820	40	884	3:00-4:00	35	948	64	1047	1931	29	0	185	0
4:00-5:00	52	886	66	1004	4:00-5:00	61	1098	74	1233	2237	57	4	208	13
5:00-6:00	34	806	57	897	5:00-6:00	35	957	74	1066	1963	33	2	173	12
TOTAL	248	4579	244	5071	TOTAL	193	5945	368	6506	11577	190	8	741	69

EASTBOUND Approach					WESTBOUND Approach					TOTAL	XING W/L		XING E/L	
Hours	Lt	Th	Rt	Total	Hours	Lt	Th	Rt	Total	E-W	Ped	Sch	Ped	Sch
9:30-10:30	121	200	68	389	9:30-10:30	4	423	44	471	860	12	0	77	1
10:30-11:30	137	184	59	380	10:30-11:30	0	310	36	346	726	36	0	53	0
11:30-12:30	161	244	98	503	11:30-12:30	0	312	45	357	860	26	0	75	0
3:00-4:00	207	445	89	741	3:00-4:00	1	231	35	267	1008	59	1	55	1
4:00-5:00	285	593	98	976	4:00-5:00	1	269	32	302	1278	65	0	130	8
5:00-6:00	356	766	97	1219	5:00-6:00	1	313	33	347	1566	52	0	110	0
TOTAL	1267	2432	509	4208	TOTAL	7	1858	225	2090	6298	250	1	500	10

(Rev Oct 06)

After extracting pedestrian volume data from these sheets, I assembled data from the built environment that I thought could possibly be significant in predicting pedestrian volume at intersections in Los Angeles. These data are public, but were assembled for a previous project at

LADOT. I looked to the literature review to inform the types of data to collect for evaluating. My built environment data (explanatory variables) include:

- Population within 0.25 mi. (`SUM_POPTTL`)
- Employment within 0.25 mi. (`EMPTOT`)
- Count of Schools within 0.25 mi. (`SCH_CT`)
- Presence of a traffic signal (`SIG`, 1 = yes, 0 = no)
- Count of transit stops within 100 ft. (`TRANSITSTOP`)
- Transit Ridership (`RIDERSHIP`)

I also dropped unneeded columns for the analysis. The final data set, which included both the extracted volumes and the built environment data, looked like the following:

```
   volume  SIG  TRANSITSTOP  RIDERSHIP  SCH_CT      EMPTOT  SUM_POPTTL
0       9  1.0          1.0        4.0       0  211.024148    7.026113
```

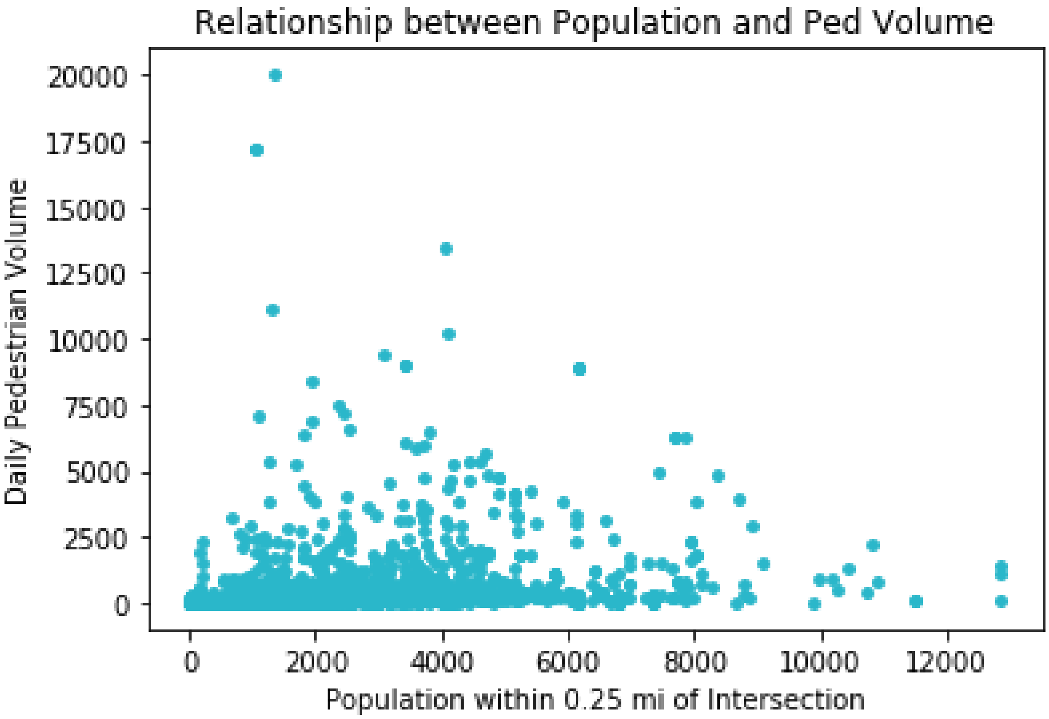
2.1 Data Cleaning

After completing the data assembly process, I inspected outlier values for the pedestrian volume counts, subsetting out those counts that are outside 3 standard deviations from the mean.

```
There are 21 volume counts outside +/- 3 st.dev from the mean volume.
```

I went back to the original PDF files to verify the count information and found 8 counts (of the 21 outliers) where the OCR failed to properly transcribe the values. I removed these 8 inaccurate counts from the dataset. Worried that there was a systematic error in my data processing, I randomly sampled 50 other volume totals. However, among those 50 sampled counts, all of them matched up with the transcribed totals, suggesting that the errors were most likely limited to those outlier totals.

During my data cleaning, I also found several volume counts that were correctly transcribed to be 0 for the all day count. This was especially odd, given that these counts sometimes occurred in dense areas of the city.



```
There are 143 counts with volume = 0
```

Further investigation yielded the discovery that occassionally LADOT requests vehicle-only counts for intersections. When this is the case, the volume is recorded as 0 for the day. Without any other method of determining when the volume at an intersection truly was 0 and when it was not part of the count, I decided to remove all instances where the volume is equal to 0.

3 Analysis

3.1 Data Exploration & Visualization

3.1.1 Target Variable (Volume)

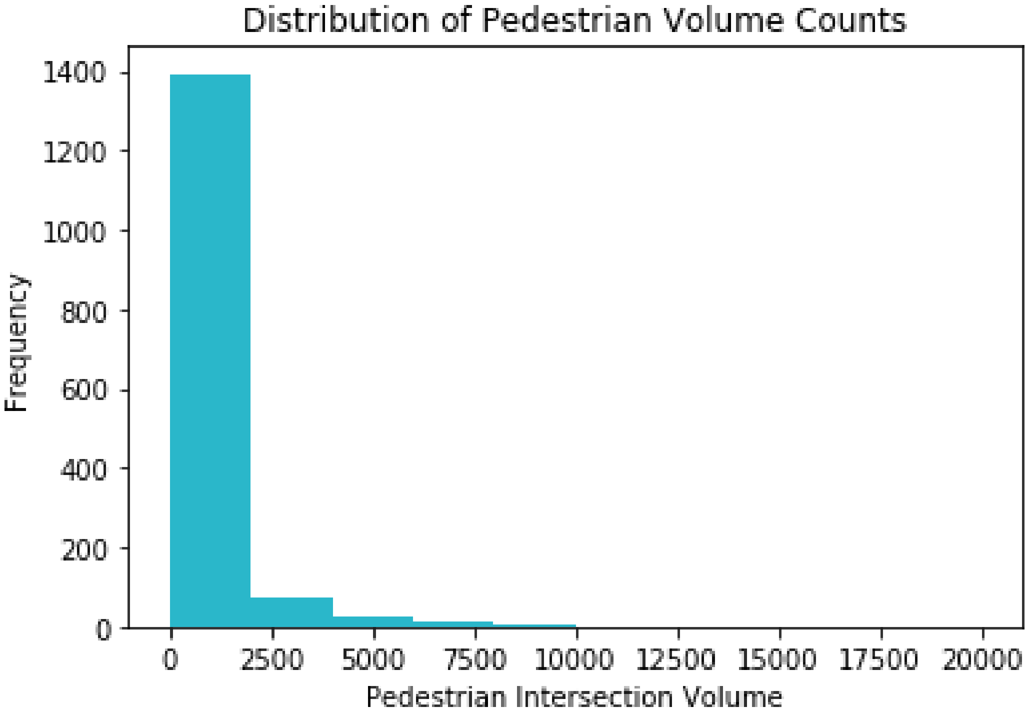
Below are the summary statistics for the target variable, weekday pedestrian volume.

```
count      1515.000000
mean        683.248185
std        1481.612605
min           1.000000
25%          85.000000
50%         229.000000
75%         593.000000
max       20005.000000
```

```
Name: volume, dtype: float64
```

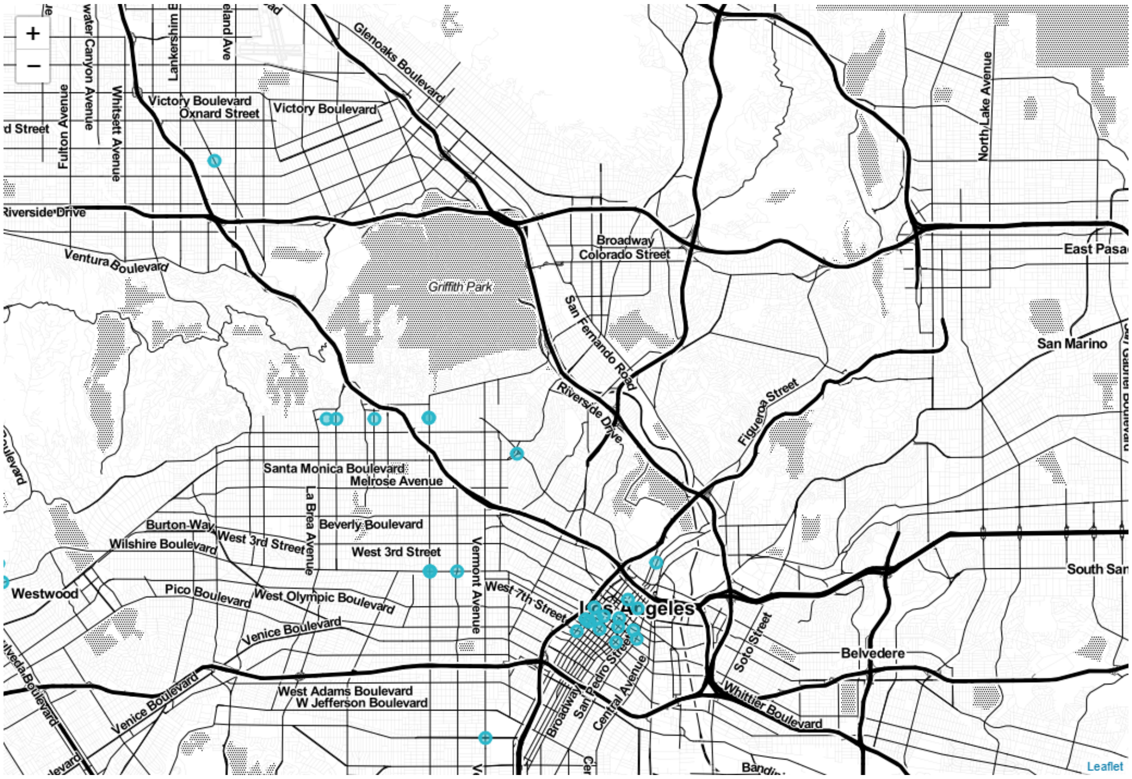
The daily number of people at an intersection in Los Angeles, according to these counts, ranges between 1 (after removing all counts equal to 0) and 20,005. Half of the intersections have a daily volume below 200; however, the mean is significantly higher (at 624), suggesting that there are some intersections with significantly higher counts. Using the same +3 standard deviation threshold as before, I can identify those intersections that appear to be outliers. This calculation resulted in a slightly different set of outliers, since this dataset excludes the erroneous counts.

```
There are 32 volume counts outside +/- 3 st.dev from the mean volume.
```



As shown in the histogram above, pedestrian volume counts has a heavy skew to the right. ALthough 75% of the counts have a daily volume below 593, some counts are as high as 20,000. This heavy skew suggests that my eventual model will take a transformed version of this variable.

Where are the intersections with the highest volume located in the City of Los Angeles? The map below shows locations with volumes outside 3 standard deviations of the mean.



Most of these locations fall within two categories:

1. Directly adjacent to a university: 4 counts are near UCLA, and 1 count is near USC
2. In a dense area: several counts are in Downtown, Hollywood, and Koreatown, some of the most dense areas in the City of Los Angeles

This suggests that density, measured in dataset by EMPTOT and SUM_POPTTL, and proximity to a school, measured in the dataset by SCH_CT, may be important features in predicting pedestrian volumes. This also suggests that these high counts, despite being outliers in the dataset, should not be discarded since they can be explained by the built environment surrounding them.

3.1.2 Features

Below are the summary statistics for the features in the dataset.

```
               SIG  TRANSITSTOP      RIDERSHIP       SCH_CT        EMPTOT  \
count  1515.000000  1515.000000     676.000000  1515.000000   1515.000000
mean      0.516172     0.937294     928.045014     0.734653   2146.467034
std       0.499903     1.321263    5937.303793     0.939846   5209.780373
min       0.000000     0.000000       0.000000     0.000000     14.084445
25%       0.000000     0.000000      43.000000     0.000000    318.393078
50%       1.000000     0.000000     129.500000     0.000000    713.423911
75%       1.000000     2.000000     511.500000     1.000000   1668.374141
max       1.000000     7.000000  104376.101400     7.000000  60793.278700
```

```
          SUM_POPTTL
count    1515.000000
mean     2678.297181
std      1877.970722
min         0.000000
25%      1303.732842
50%      2284.940640
75%      3557.542856
max     12858.609440
```

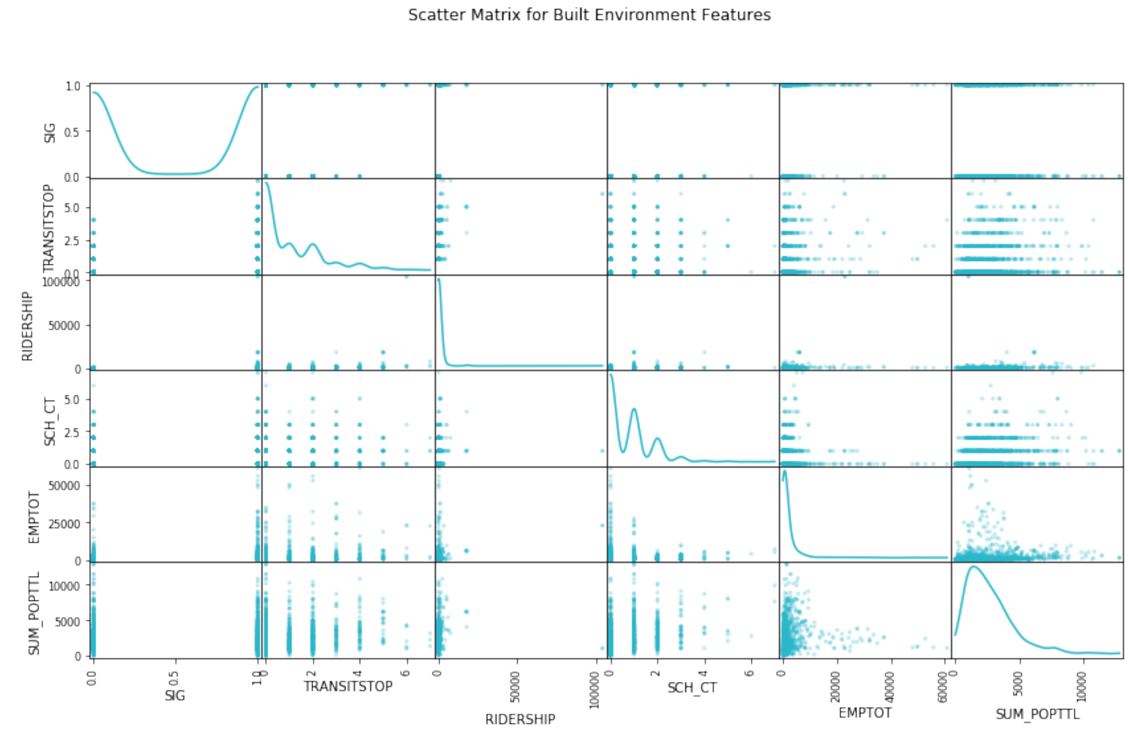
The feature SIG, presence of a signal, takes on a value of 0 (not present) or 1 (present). The mean value of this feature is .51, indicating that more than half the intersections in the dataset are signalized. The second feature TRANSITSTOP, counts the number of transit stops within 100 ft. of the intersection. In many cases, there are no transit stops at the intersection being measured (resulting in a value of 0), but in at least one case there is 7 different transit stops at a single intersection.

The feature RIDERSHIP measures the weekly transit ridership at the intersection. This feature is not normally distributed, with the median transit ridership at an intersection equal to 129 but the mean ridership at 928, suggesting a heavy right skew. Since the transit ridership first depends on a transit stop existing, there were several rows where RIDERSHIP has no value (resulting in a row count equal to 676). Since this was the case if a transit stop did not exist, these NA values needed to be replaced with the value of 0 in the data preprocessing process.

The feature SCH_CT measures the number of schools (elementary, middle, high, university) within 0.25 mi. of the intersection. In most cases, there is not a single school that close to an instersection; however, at one intersection there is 7 schools nearby.

The features EMPTOT and SUM_POPTTL measure the density of the area surrounding the intersection. Both are derived from the latest American Communities Survey estimates for people living (SUM_POPTTL) and working (EMPTOT) within 0.25 mi. of the intersection. The summary statistcs above suggest that these variables are distributed with a right-skew and may also need to be transformed to achieve a normal distribution.

The scatter matrix below shows the relationship between each of the feature pairs as well as the kernel density estimate for each feature in the diagonal.



The kernel density estimates confirm that two of the features, `EMPTTOT` and `RIDERSHIP`, are not normally distributed and exhibit a heavy right-skew. In order to prevent the very large values from negatively affecting the performance of the algorithm, these features will need to be rescaled during the data preprocessing step. The only qualitative variable, `SIG`, is already coded to the values equal to either `0` (intersection does not contain a signal) or `1` (intersection does contain a signal).

3.2 Algorithms and Techniques

3.2.1 Linear Regression

My proposed solution is to build a regression model that can take inputs from the built environment and predict the daily volume of pedestrians at an intersection. For this project, it is not just important to be able to accurately predict the pedestrian volume; it is also important to understand how the characteristics of the built environment affect the volume. A linear regression model produces coefficients that are easy to interpret. In fact, my choice of built-environment features to measure is informed by linear models from previous research. I anticipate my resulting model to take one of two forms shown below:

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ...\beta_j X_{ji}$

$Y_i = exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ...\beta_j X_{ji})$

where:

Yi = weekday pedestrian volume at intersection

X_{ji} = value of explanatory variable j at intersection i

β_j = model coefficient for variable j

3.2.2 Decision Tree Regression

In addition to linear regression, I will also build a decision tree regressor for comparison. A decision tree regressor will create a set of splitting rules for the feature data to then segment the predictor space into a number of simple regions.

3.2.3 Benchmark

There have been a few attempts to build models estimating the effect of the built environment on pedestrian traffic volume, but none have focused on Los Angeles. Most have used a sample set smaller than my own, and almost none reported a test R^2 alongside the train R^2. Below is a table with the results from two of these models.

Study	Model Structure	Sample Size (n)	Train R^2 Value	F-Value	Test R^2 Value
SF (1)	Linear	50	0.804 Adj.	34.4	0.387
SD (5)	Linear	79	0.516 Adj.	24.112	None Reported

The San Diego model did not perform a test R^2, so it is not possible to determine how well the model generalizes. Similarly, the San Francisco model did not initially build a train / test procedure into the modelling process. Instead, after the model was complete, the researchers validated the model with 49 pedestrian counts from a prior study. The correlation between the predicted and actual volumes at those intersections was 0.387. This R^2 is the benchmark score for comparing the results of this study.

4 Methodology

4.0.1 Data Preprocessing

Most of the data preprocessing for this project is documented in Data Assembly. After assembling the data, the summary statistics revealed that there were several `NA` values for the `RIDERSHIP` feature, which meant that there were no transit stops within 100ft of the intersection. For the purposes of this study, I can interpret this as a `RIDERSHIP` value of 0, so I can fill all the `NA` values with a value of 0 for the analysis.

4.0.2 Transforming Skewed Continuous Variables

As also noted in the data exploration, the target variable `volume` and the features `RIDERSHIP` and `EMPTOT` are not normally distributed, all having heavy right skewed distributions. In order to prevent the very large values from negatively affecting the performance of the algorithm, I rescaled them using the natural logarithm.

5 Implementation

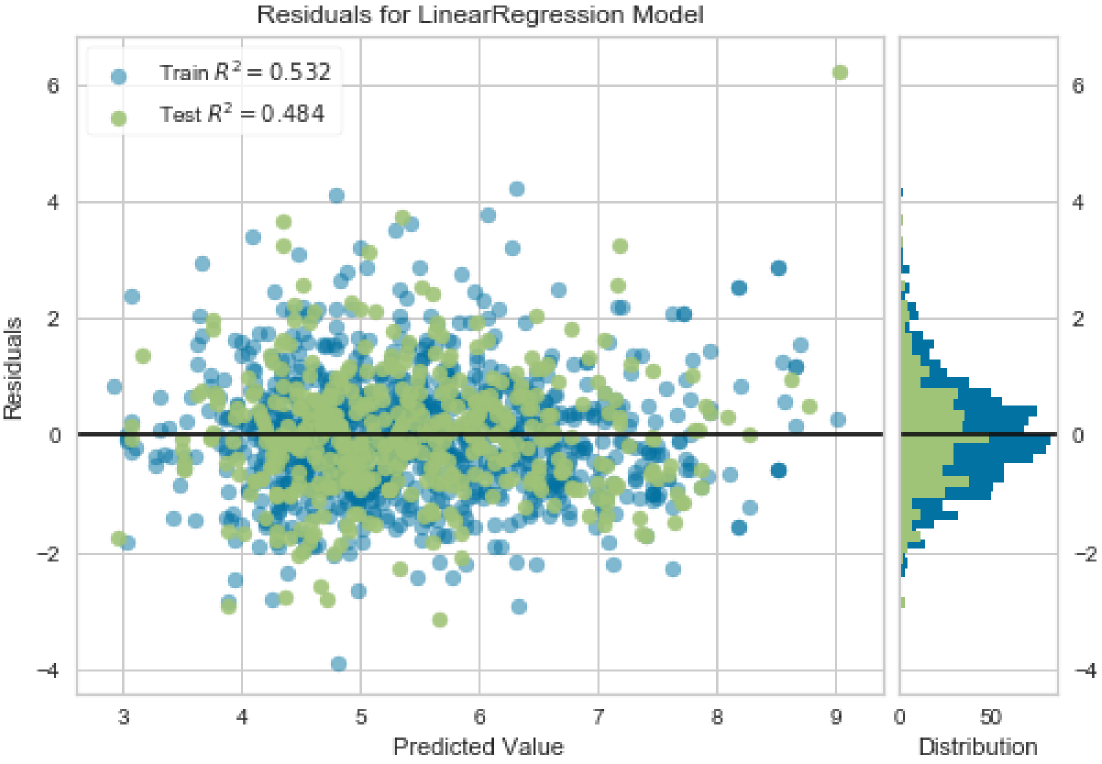
In order to assess the true predictive power for each of the models, I split the 1,511 samples into a training and testing set, with 30 percent of the samples held out for testing and the remaining 70 percent used for training and model validation. The initial models included an ordinary least squares (OLS) regression model and a decision trees regression model with a maximum tree depth of five.

```
Training set has 1060 samples.
Testing set has 455 samples.
```

5.1 Refinement

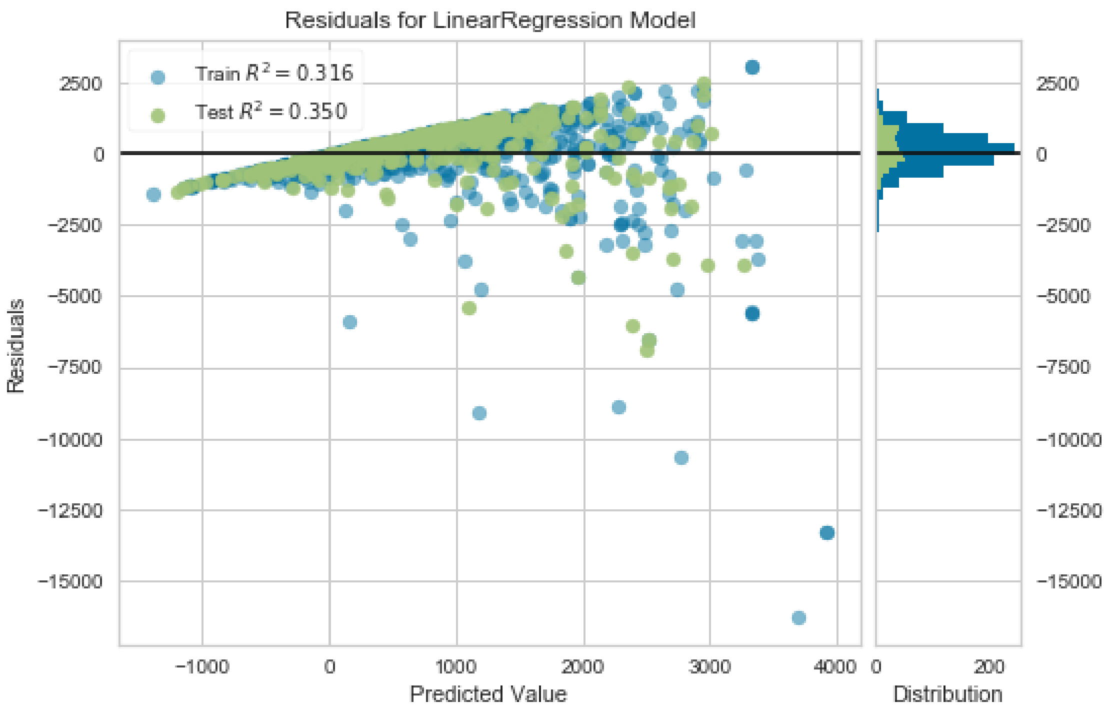
5.1.1 Linear Regression Model

The linear regression model produces an R^2 value of 0.484 on the test set, indicating that the model can explain 48% of the variation in log-transformed daily pedestrian volume from the givien features. Both figures below suggest that current linear model is a good fit for the data. The figure below and to the left, a plot of the residuls against the predicted values for the log of the daily pedestrian volumes, shows no discernable pattern. The figure below and to the right shows the residuals to also be normally distributed around 0.



```
Mean squared error: 1.12
```

The figure below shows the results of a regresssion against the non-transformed version pedestrian volume. The evident pattern in the residuals suggests that a linear model with the non-transformed volume data is not a good fit. In additon, the model also shows much lower R^2 values compared to the model with the log-transformed volumes above.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.532
Model:                            OLS   Adj. R-squared:                  0.529
Method:                 Least Squares   F-statistic:                     199.5
Date:                Sat, 22 Sep 2018   Prob (F-statistic):          9.19e-170
Time:                        18:50:44   Log-Likelihood:                -1519.1
No. Observations:                1060   AIC:                             3052.
Df Residuals:                    1053   BIC:                             3087.
Df Model:                           6
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const           1.6135      0.163      9.869      0.000       1.293       1.934
SIG             0.6244      0.072      8.724      0.000       0.484       0.765
TRANSITSTOP    -0.0243      0.045     -0.542      0.588      -0.112       0.064
lgRIDERSHIP     0.1748      0.023      7.683      0.000       0.130       0.220
SCH_CT          0.0782      0.034      2.268      0.024       0.011       0.146
lgEMPTOT        0.3758      0.025     15.030      0.000       0.327       0.425
SUM_POPTTL      0.0002    1.8e-05     12.006      0.000       0.000       0.000
==============================================================================
Omnibus:                       63.598   Durbin-Watson:                   2.119
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              101.268
```

```
Skew:                          -0.468   Prob(JB):                     1.02e-22
Kurtosis:                       4.191   Cond. No.                     1.74e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.74e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The table above shows the coefficients for each of the features in the linear model. All of the features, with the exception of `TRANSITSTOP` are shown to have significant P Values. In addition to not being signifcant in predictor value (with a P Value of .588), the model suggests a relationship with `TRANSITSTOP` that does not make sense. The negative coefficient suggests that pedestrian volume decreases with an increase in the number of transit stops. I removed this feature from future models.

```
OLS Test R^2 value: 0.48
OLS Mean squared error: 1.12
```

5.1.2 Decision Trees

I re-fit my decision tree model with the updated test/train set (without the `RIDERSHIP` feature). This regressor yielded the similar results as my linear model, with a test R^2 of 0.47. Decision Tree algorithms allow for several parameters of the model to be tuned to improve predictive performance. In additon to tuning parameters within the model, I decided to add gradient boosting.

```
DT Test R^2 value: 0.47
DT Mean squared error: 1.15
```

5.1.3 Improving Decision Trees with Gradient Boosting

To improve the predictive performance of my Decision Tree model, I added gradient boosting, a technique in ensemble learning where new models are added sequentially to correct for the errors in previous models. Gradient boosting reduces these errors through a gradient descent algorithm. Models stop being created when no further improvements can be made; these models are then combined to create a final prediction. Specifically, I decided to use the XGBoost algorithm, which is one of the most popular algorithms among Kaggle Winning Solutions. After performing a grid search on possible parameters, the best performing XGB Model contained the following parameters:

```
C:\Users\Tim\Anaconda3\lib\site-packages\sklearn\cross_validation.py:41: DeprecationWarning: Th
  "This module will be removed in 0.20.", DeprecationWarning)


XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
       colsample_bytree=0.4, gamma=0.5, learning_rate=0.1,
       max_delta_step=0, max_depth=5, min_child_weight=6, missing=None,
       n_estimators=1000, n_jobs=1, nthread=None, objective='reg:linear',
       random_state=0, reg_alpha=10, reg_lambda=1, scale_pos_weight=1,
       seed=None, silent=True, subsample=1)
XGB Test R^2 value: 0.56
XGB Mean squared error: 0.95
```

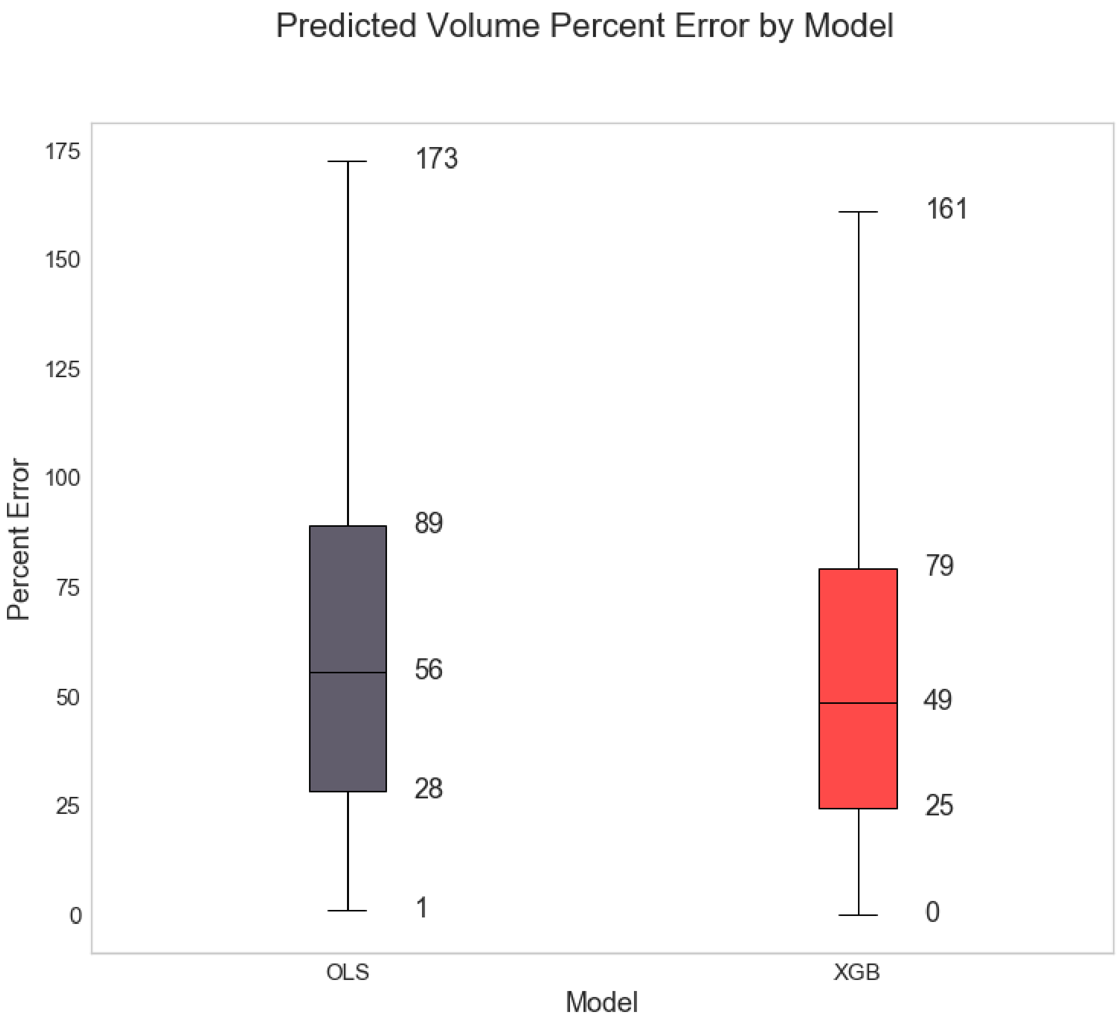
6 Results

6.1 Model Evaluation and Validation

Model Structure	Test R^2 Value	MSE
Linear, Ordinary Least Squares	0.48	1.12
Decision Trees	0.47	1.15
Decision Trees (Boosted)	0.56	0.95

The table above shows the results from my three models. My initial linear OLS model has test R^2 of 0.48, meaning that the linear model can explain 48% of the variation of pedestrian volume at intersections. I was able to improve this R^2 to .55 using the Boosted Decision Tree model. Although not my primary metric (since it is not comparable across studies), the mean squared error (MSE) was also lowest for the Boosted Decision Tree model as well, further justifying its selection as the best model for prediction. Since the testing set included randomly selected 455 samples across Los Angeles among the counts, I have a high amount of confidence in my Test R^2. Given that the counts were also distributed fairly randomly throughout Los Angeles, I believe my model would generalize well for other counts in the city.

The boxplot below shows the distribution of percent error for predicted volumes in the test set for the linear and boosted decision tree models.



The boxplot above shows that half of my predicted values had an error of less than 50%. For exploratory purposes and predictive purposes, I believe this to be an acceptable outcome and inline with similar research. The SF model, for example, noted that "there were noticable differences (more than 50%) between the model volumes and count volumes at a majority of intersections" (1).

The boxplot above does not include outliers.

```
There are 61 predictions with an error greater than 161 percent of the true value.
```

6.1.1 Interpreting the Linear Model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.532
Model:                            OLS   Adj. R-squared:                  0.530
Method:                 Least Squares   F-statistic:                     239.5
```

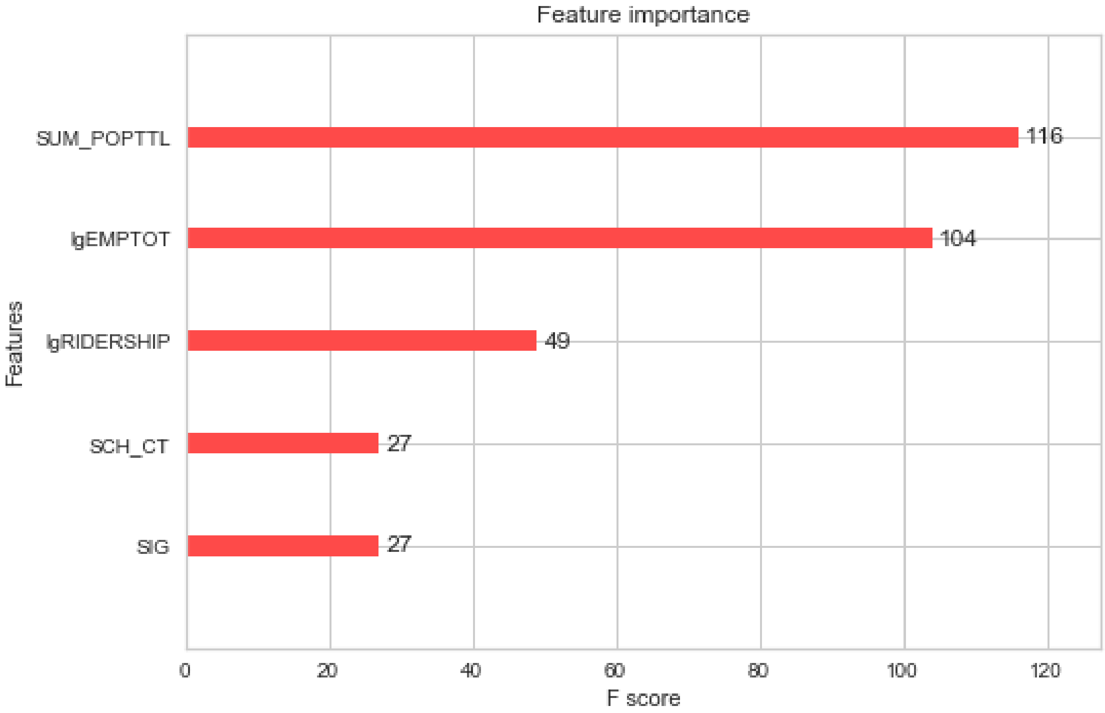
```
Date:                Sat, 22 Sep 2018   Prob (F-statistic):          6.54e-171
Time:                        18:50:46   Log-Likelihood:                -1519.2
No. Observations:                1060   AIC:                             3050.
Df Residuals:                    1054   BIC:                             3080.
Df Model:                           5
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const           1.6140      0.163      9.876      0.000       1.293       1.935
SIG             0.6251      0.072      8.737      0.000       0.485       0.765
lgRIDERSHIP     0.1648      0.013     12.584      0.000       0.139       0.190
SCH_CT          0.0784      0.034      2.273      0.023       0.011       0.146
lgEMPTOT        0.3757      0.025     15.030      0.000       0.327       0.425
SUM_POPTTL      0.0002    1.8e-05     12.004      0.000       0.000       0.000
==============================================================================
Omnibus:                       62.871   Durbin-Watson:                   2.119
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               99.760
Skew:                          -0.464   Prob(JB):                     2.18e-22
Kurtosis:                       4.181   Cond. No.                     1.74e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.74e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

The table above shows the effect of each feature on the target variable (log-transformed volume) when holding all other features constant. * Presence of a Signal: We can expect signalized intersections to have pedestrian volumes that are 87 percent higher compared to those that are not signalized, since exp(.6251) = 1.87. * Transit Ridership: Both the dependent variable (pedestrian volume) and independent variable (transit ridership) are log transformed, so they will exhibit an elastic relationship. A one percent increase in transit ridership corresponds to a .16 percent increase in the pedestrian volume at the intersection. * School Count: Each additional school within 0.25 mi. of the intersection results in an increase in the pedestrian volume by 8.16 percent, since exp(.0784) = 1.0816. * Employment: Both the dependent variable (pedestrian volume) and independent variable (employment) are log transformed, so they will exhibit an elastic relationship. A one percent increase in employment within 0.25 mi. of the intersection yields a .38 percent increase in the pedestrian volume at the intersection. * Population: A one unit increase in the population within 0.25 mi. of the intersection results in an increase of the pedestrian volume by .02 percent, since exp(.0002) = 1.0002.

6.1.2 Interpreting the XGBoost Model

Boosted decision trees provide a score that indicates how useful a feature is within the model. Features that are used more often to make key decisions within decision trees are scored as having a higher relative importance. For a single tree, the score is calculated from the amount that each attribute split point improves the performance measure (in this case, the R^2 value), weighted by the number of observations the node is responsible for. For boosted trees, the feature importance is the average of all the feature scores among the decision trees in the model.



In this model, employment and population are ranked the highest in terms of feature importance for predciting pedestrian volume. Transit ridership is also scored to be relatively important in my boosted trees model, while school count and signalized status are far less important in predicting pedestrian volume.

6.2 Justification

Based on the available literature, my results were in-line with my expectations. As discussed in the Benchmark section, it is difficult to compare my results with many of the other models that have been completed, since most did not create a train / test split on the data to begin with. Both of my models outperformed the San Francisco model, which had a higher Adjusted R^2, but had a worse correlation between predicted and actual volumes on data that was unseen by the model (test R^2).

Study	Model	Test R^2 Value
SF (1)	Linear, Ordinary Least Squares	0.39
This work	Linear, Ordinary Least Squares	0.48
This work	Decision Trees (Boosted)	0.56

Both of my final models adress the two parts of my original problem statement:

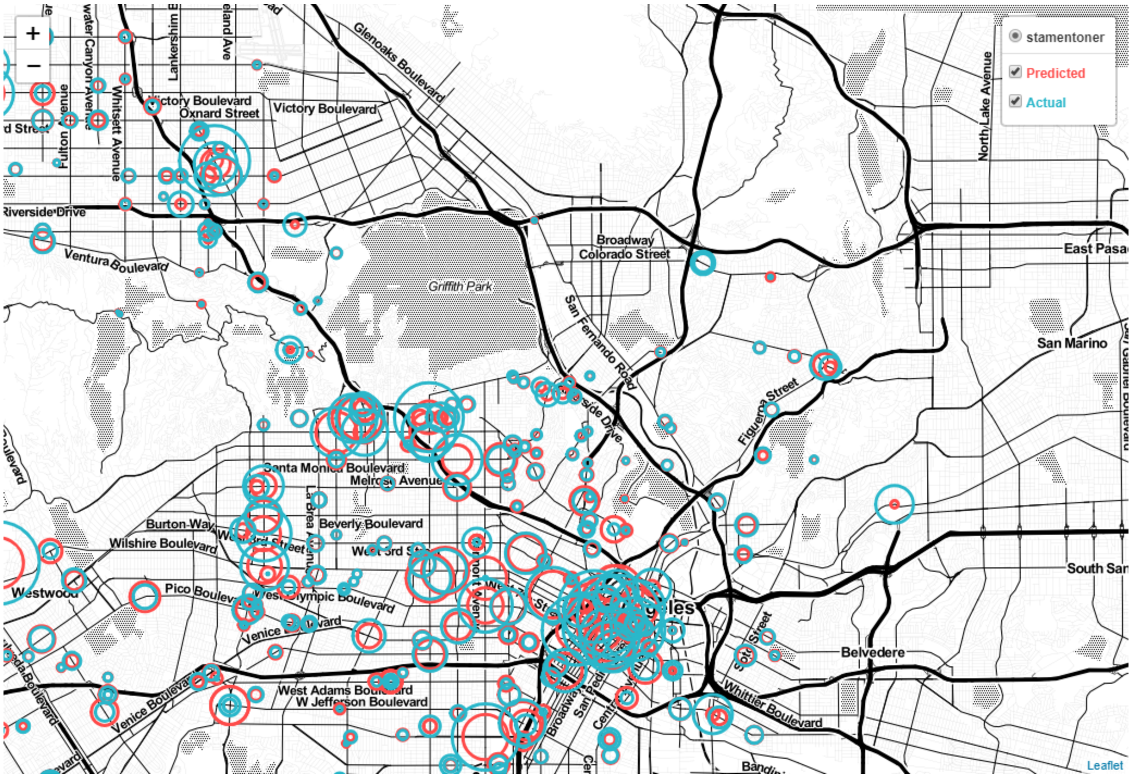
1. Understanding the affect of features: The linear model produced a an interpretable summary that explains how each of the features will affect the pedestrian volume, holding all other features constant.
2. Prediction: The XGB model is the best predictor with the ability to explain 56 percent of the variation in log-transfomed daily intersection ped volume. After transforming the volumes back to the original counts, half of the predicted values within the test set had an error of less than 50% of the true value.

7 V. Conclusion

7.1 Free-Form Visualization

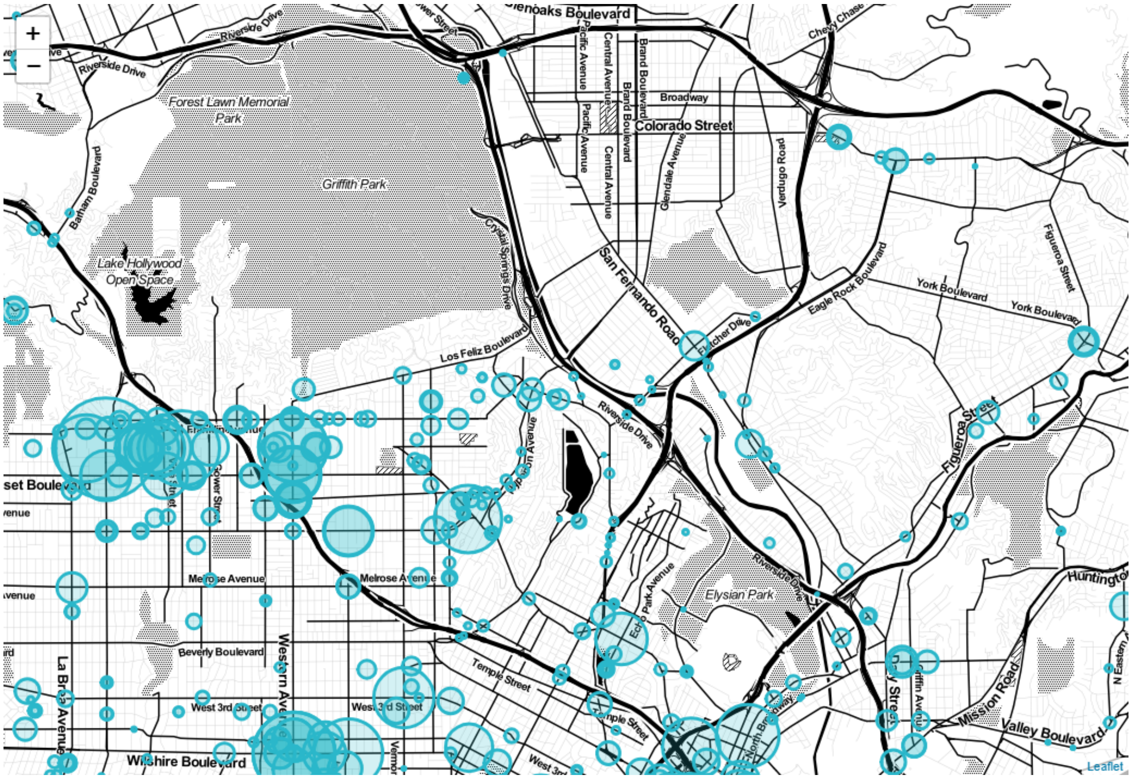
7.1.1 Visualizing the Test Results

The map below shows the difference between the predicted and actual values for the pedestrian volume in the test set. The predicted volumes are from the XGB model.



7.1.2 Visualizing the Entire Dataset of Pedestrian Volumes

Below is a map of all the pedestrian volume counts in Los Angeles used in this analysis, symbolizing the pedestrian volume at each intersection by a circle with the area equal to the volume. The map below supports the general takeaway that there are higher pedestrian volumes near downtown, Hollywood, and koreatown. Volume counts are generally lower in lower-density places such as the San Fernando Valley.



7.2 Reflection

For this project, I was able to take data that is publicly available to build a model that predicts the daily amount of people at an intersection in the City of Los Angeles, employing two widely used methods for regression: ordinary least squares linear regression and decision trees regression. In addition, I was able to improve the predictive power of my decision tree model by employing gradient boosting.

The most challenging and time-consuming aspect of the project was the preprocessing of the data. I spent a considerable amount of time familiarizing myself with the OpenCV library so that I could employ it on this task. Before implementing the OpenCV / OCR solution to process the data, I originally tried to read the data using PDF text extraction tools; however, the performance using those tools was quite poor, and I didn't feel that it would be relaible enough for this project. The OpenCV library method produced more accurate totals, but could only process a smaller set of the PDFs.

The results of this project were generally in line with results similar efforts elsewhere. As discussed previously, it is difficult to compare results with other efforts that did not perform a train / test methodology, and it is likely that those efforts created models that overfit the training data and therefore would not generalize well. Because of this difference in methodologies, and because the much larger sample size in my study, I feel confident that my models would perform better in a more generalized setting.

7.3 Improvement

One improvement that I could have made earlier on would be to rethink the methodology for transforming at least one of the variables, SCH_CT. During the initial data collection, I aggregated across all school types; however, I identified the presence of a university as one of the two factors that seemed to explain the presence of outlier values. It may be the case that the presence of a university affects pedestrian volumes differently than other schools, and should be its own separate variable.

Generally, given the data that I had to build the model, I feel relatively confident that my model for predcition, the XGBoost Decision Tree algorithm, is likely the best model for prediction. However, there are additional prediction techniques that were not evaluated in this study, such as K-Nearest Neighbors.

Despite having a larger dataset compared to previous similar efforts, this modeling solution could be improved further with more data on the target variable, daily pedestrian volume. The OpenCV / OCR solution was limited to the approximately 1,600 counts; however, there are least 5,000 additional sheets that could not be processed in this effort. Obtaining the volume counts from those sheets this would increase the amount of training data by roughly 4 times. Beyond what I will be able to get from these historical counts, the field of machine learning has advanced to the point where we can now install cameras and get much larger datasets on ped volume counts. The City of Los Angeles has not yet implemented this strategy for collecting pedestrian counts; however, it is only a matter of time before that becomes standard in many cities. When this does happen, we will have much more

In addition to more data on the target variable, it is highly likely that a better model could be constructed using additional, different data from the built environment. I was able to assemble the data that the literature has shown to be the most important in terms of predicting pedestrian volume; however, I was not able to get the range of data that other researchers have examined. It is also possible that there is some aspect of the built environment that is important but has not yet been revealed in the research.

8 References

1. Schneider, R. J., T. Henry, M. F. Mitman, L. Stonehill, and J. Koehler. Development and Application of a Pedestrian Volume Model in San Francisco. Transportation Research Record: Journal of the Transportation Research Board, No. 2299, 2012, pp. 65–78.
2. Liu, X., and J. Griswold. Pedestrian Volume Modeling: A Case Study of San Francisco. Association of Pacific Coast Geographers Yearbook, Vol. 71, 2009.
3. Pulugurtha, S. S., and S. R. Repaka. Assessment of Models to Measure Pedestrian Activity at Signalized Intersections. In Transportation Research Record: Journal of the Transportation Research Board, No. 2073, Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 39-48.
4. Schneider, R. J., L. S. Arnold, and D. R. Ragland. Pilot Model for Estimating Pedestrian Intersection Crossing Volumes. In Transportation Research Record: Journal of the Transportation Research Board, No. 2140, Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 13-26.
5. Jones, M. G., S. Ryan, J. Donlan, L. Ledbetter, L. Arnold, and D. Ragland. Seamless Travel: Measuring Bicycle and Pedestrian Activity in San Diego County and Its Relationship to Land Use, Transportation, Safety, and Facility Type. Alta Planning and Design and Safe Transportation Research and Education Center, University of California, Berkeley, 2010.

6. Haynes, M., and S. Andrzejewski. GIS Based Bicycle & Pedestrian Demand Forecasting Techniques. Presentation to Travel Model Improvement Program, U.S. Department of Transportation. Fehr & Peers Transportation Consultants, San Francisco, Calif., 2010.
7. Miranda-Moreno, L. F., and D. Fernandes. Modeling of Pedestrian Activity at Signalized Intersections: Land Use, Urban Form, Weather, and Spatiotemporal Patterns. In Transportation Research Record: Journal of the Transportation Research Board, No. 2264, Transportation Research Board of the National Academies, Washington, D.C., 2011, pp. 74-82.