Predicting the Optimum type and location of a restaurant based on demographics

Introduction:

Establishing a small business in any location is tough and its success is dependent on various factors. But the location of a place greatly determines the success of a business. It is affected by various parameters with the demographics and the type of location affecting the business. Every Neighborhood is different and has a different demographic which could indicate the number of businesses the neighborhood could sustain.

Based on common logic, we could make several assumptions:

Richer neighborhoods have more expensive businesses because of the purchasing power of the people in the neighborhood.

The age of the people in a neighborhood could affect their preferences and tastes.

The cost of renting and land could affect the feasibility of running a business in the area.

Such common assumptions could give an insight, but it cannot show the direct effect of such features on the businesses. This project aims to analyze and explore **the effect of key demographic factors on the type and number of Restaurants in a neighborhood**.

The project would be addressed to people who are going to start a new Restaurant in a location and would like to know the feasibility of establishing a new Restaurant and what type of cuisine or type the restaurant should belong to.

Data Description:

The model takes into consideration the **five Boroughs of New York city** and uses the neighborhoods based on the Neighborhood Tabulation Areas (NTAs) which are aggregations of the Census tracts of New York.

For our model, the following Data will be used as prediction parameters and all the parameters are chosen on a per neighborhood basis:

To find the size of customer base:

- **Area Population density**
- **Daytime Population** - The population of an area during working hours which is commuter-adjusted

Understand the characteristics of the customers:

- **Median Income** of the resident population
- **Median Age** of the Population

To find the quality of neighborhood and thoroughfare:

- **Median Rent**
- **Median Property Value**
- **Landmarks and attractions** within the Area
- **Average Daily Traffic** in the Area

These features are all available open-source from the Government of USA and New York State. The Data has been collected through the following channels:

- Demographic Data and Income and Housing Data was obtained using the Census Data Portal.
- The Geographical Data of the neighborhoods was obtained from New York City Open Datasets. The shapefiles provided give the Name, Area Code, Location and shape of the NTA Areas.
- The Data about traffic was obtained from New York Dept. of Transportation.
- The data about landmarks in the neighborhoods and the actual Data about the types and number of restaurants in each neighborhood was obtained from the Foursquare API.

Data Processing:

The initial Data available from the demographic Datasets are stored as CSV files which are later read using Pandas and the required columns are only read. The directly obtained data columns are as follows: **Population, Density, Median Age, Median Income, Median Rent and Median Land Value**. Using these initial data, further Data like **Daytime population** is calculated using the population multipliers and **Average Annual Traffic** is obtained from the individual road traffics and merged into a single Dataframe. The number of people in every level of income are also categorized to represent each neighborhood.

The Data collected from Foursquare are about each venue in the following Categories:

- Restaurants
- Coffee Shops
- Drinks
- Historical Sites, Art Venues and Landmarks

The first three categories are used as the target variables to be predicted and the Arts and attractions Data is used as a feature.

Further Data is obtained about landmarks from Foursquare to find the popularity of each place. The number of photos, tips and likes each place has is used to find the actual number of visitors and a crude regression line is used to form a new variable called **Cultural Factor** which takes into account the category and visitors of each place in a neighborhood has and aggregates them per neighborhood.

After the initial processing, the Data consisted of 195 rows each representing one NTA area and 15 features as columns. The areas representing Parks, Cemeteries and Airports and other government property was dropped because they have no demographic data.

Methodology:

Aim:

The overall aim of the analysis is to use machine learning models to learn and predict the number of occurrences of each target in a particular area based on the demographic features. The targets are then compared with the actual counts of the restaurants and it can be used to find how many more similar restaurants can be supported in the particular demographic. The locations with the highest predictions are the areas of interest which can be subject to direct human analysis.

Targets:

The targets were initially processed from the categories of restaurants obtained individually. Various separate categories were merged with similar ones to create classes of restaurants which will be predicted.

Pizza Place	1288	Coffee Shop	1103	Bar	2330
Deli / Bodega	1148	Café	821	Lounge	333
Chinese Restaurant	937	Donut Shop	533	Nightclub	180
Asian Restaurant	705	Bakery	131	American Restaurant	129
Italian Restaurant	698	Tea Room	100	Italian Restaurant	72
Mexican Restaurant	627	Dessert Shop	49	Asian Restaurant	69
Bakery	582	Sandwich Place	40	Gastropub	60
Japanese Restaurant	555	American Restaurant	33	Brewery	53
American Restaurant	500	Bar	31	Mexican Restaurant	45
Sandwich Place	496	Diner	31	Restaurant	36
Name: category, dtype: int64		Name: category, dtype: int64		Name: category, dtype: int64	

The Top 10 Categories of obtained under section: Restaurants, Coffee Shops, Drinks.

All these individual categories were aggregated and categories with very low counts were dropped from the analysis. The individual restaurant Data was aggregated using Groupby over each neighborhood and the number of restaurants of each type were represented as integers under columns showing the category names.

These processes finally led to 15 different features and 14 types of targets. The feature selection and PCA steps were done dynamically during the learning process to identify their effects on the results. The features and results obtained are as follows:

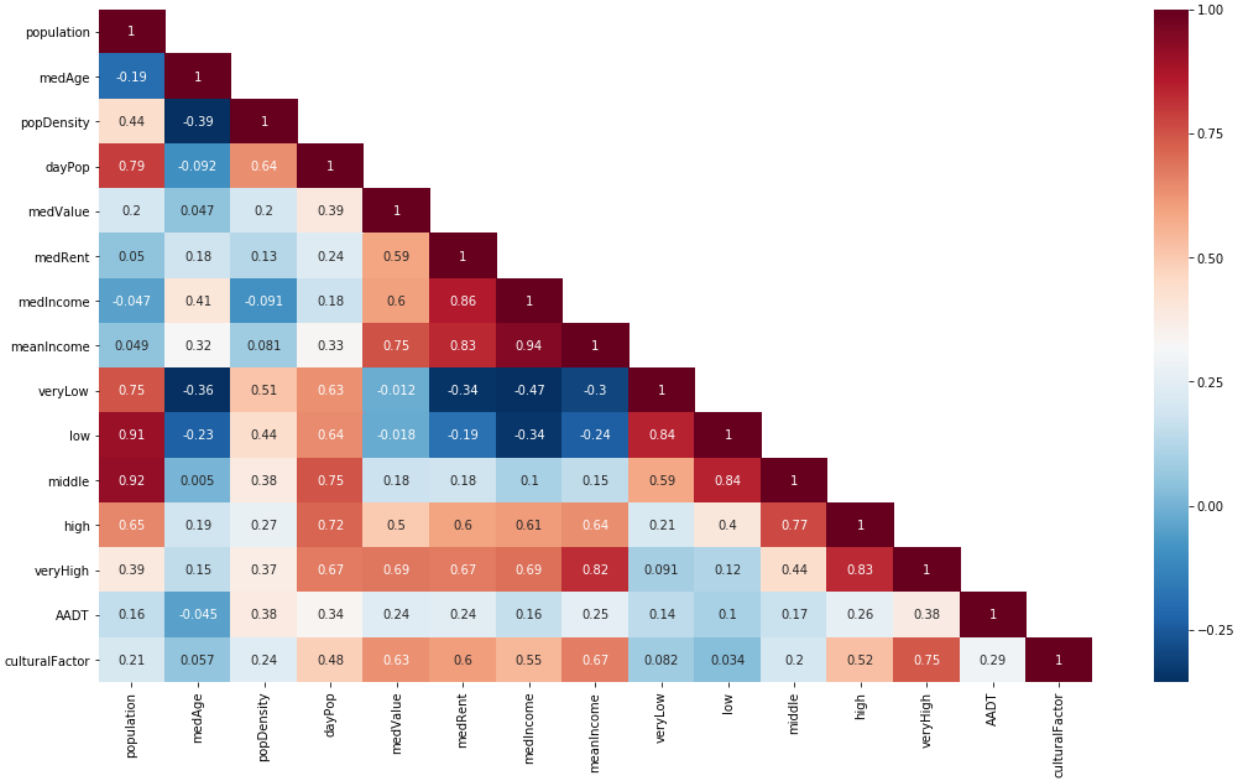
population	medAge	popDensity	dayPop	medValue	medRent	medIncome	meanIncome	veryLow	low	middle	high	veryHigh	AADT	culturalFacto
24212.0	37.1	14988.0	21307.0	856535.0	2278.0	125817.0	205275.0	1279.0	1201.0	2008.0	3355.0	3272.0	11051.190476	24.12541
67681.0	43.9	6625.0	59559.0	476965.0	1180.0	57150.0	79613.0	6637.0	5298.0	6797.0	5785.0	1633.0	12013.800000	16.60875
35811.0	44.3	12927.0	31514.0	561046.0	1194.0	36802.0	63703.0	5762.0	2697.0	3301.0	2006.0	791.0	12252.733333	3.00000
31132.0	39.0	4986.0	27396.0	457834.0	676.0	27345.0	49358.0	5381.0	2285.0	2158.0	1115.0	297.0	15428.300000	38.95058
16436.0	58.0	11656.0	14464.0	311186.0	905.0	40316.0	58752.0	3169.0	1790.0	2212.0	955.0	275.0	8256.600000	0.00000

	American Restaurant	Asian Restaurant	Bakery/Dessert	Bar	Chinese Restaurant	Coffee Shop	Deli / Bodega	Diner	European Restaurant	Fast Food Restaurant	Japanese Restaurant	Latin American Restaurant	Nightclub/Lounge	Pizza Place
0	2	4	12	7	3	12	8	5	6	9	6	4	2	6
1	5	1	16	13	3	15	7	10	11	17	15	1	3	10
2	1	2	10	3	1	12	2	4	6	5	5	1	2	2
3	4	2	6	6	1	5	6	5	1	16	2	10	3	10
4	1	0	2	3	1	2	4	5	1	6	1	0	1	4

Features Dataframe (Top) and Targets Dataframe(Bottom)

Exploratory Data Analysis:

In order to explore the relationships between the Features and targets and between themselves, heatmaps are created to explore their correlations and understand the causations. From this heat map, it is observed that some features have a high correlation with each other and could lead to multicollinearity and thus bias the results. So, these features have to be dropped or treated. Here it was chosen to employ PCA to reduce the dimensionality of collinear features.

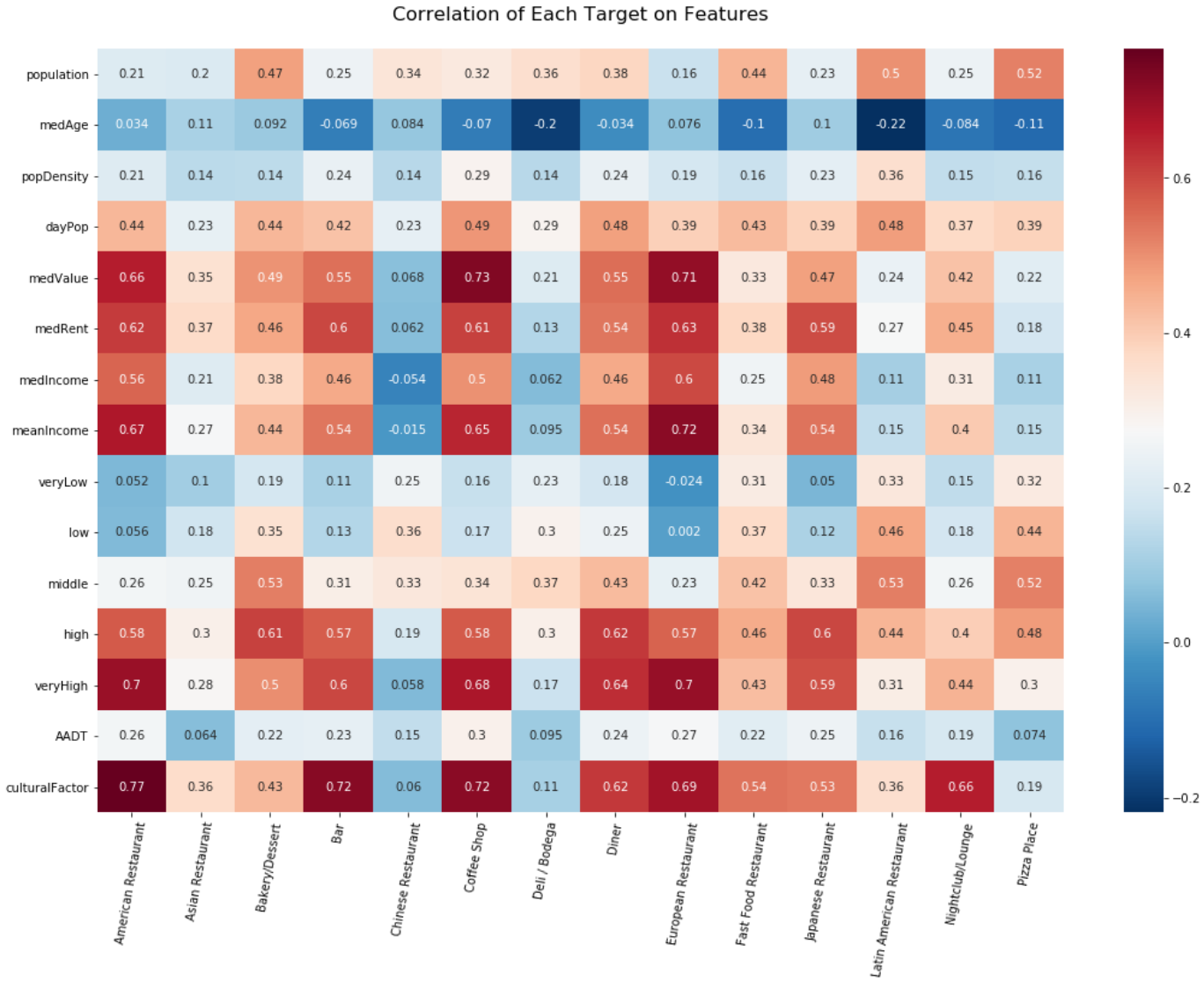


Heatmap of correlations of Feature variables with themselves

These inferences are made based on the heatmap between the targets and the features which is shown below:

- Population based demographic features influence the targets only by a small factor. Population and daytime Population affects the presence of Fast-Food Restaurants, Bakeries, Coffee Shops and Pizza Places the most, which implies they are usually takeaways or small Snacks.
- The median Age of an area generally has a slight negative correlation with the overall presence of restaurants which indicates that younger the people are, higher their tendency to eat out and hence higher support for restaurants

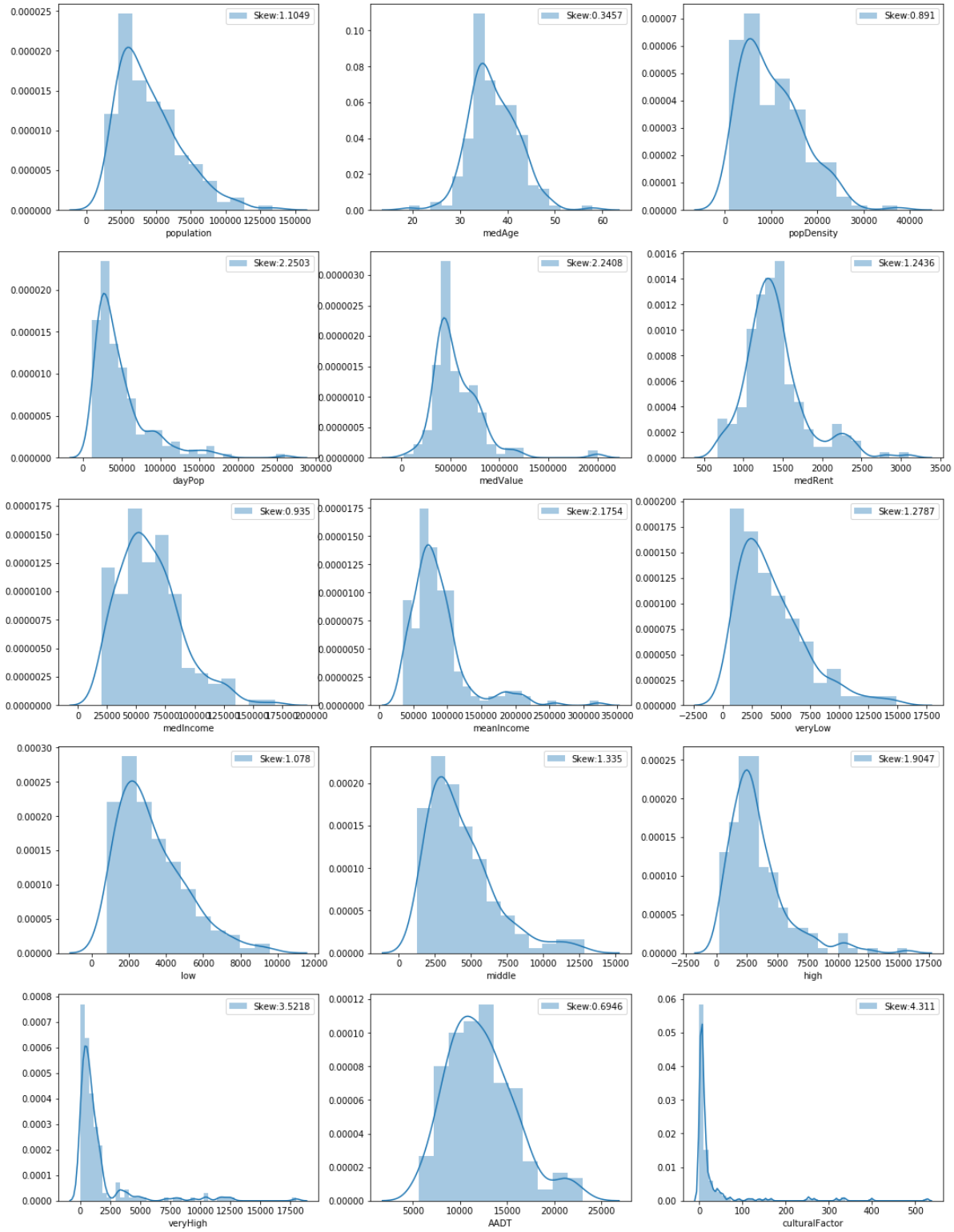
- The heatmap gives us an outline of how the people tend to eat based on their affluence. It is seen that people with higher income eat in high quality American restaurants and European Restaurants. It is also seen that Japanese restaurants are more familiar to people with higher income.
- It is seen that the presence of Landmarks and Tourist Attractions greatly determine the presence of Bars and Coffee Shops.
- The high correlation between High income neighborhoods and Landmarks could be attributed to the fact that their presence simply makes for better neighborhoods and attracts higher income people.
- Asian Restaurants, Delis and Pizza places are not highly correlated to any feature which indicates that they are found widespread and more or less evenly distributed over all neighborhoods.



Heatmap of correlations of Features vs Targets

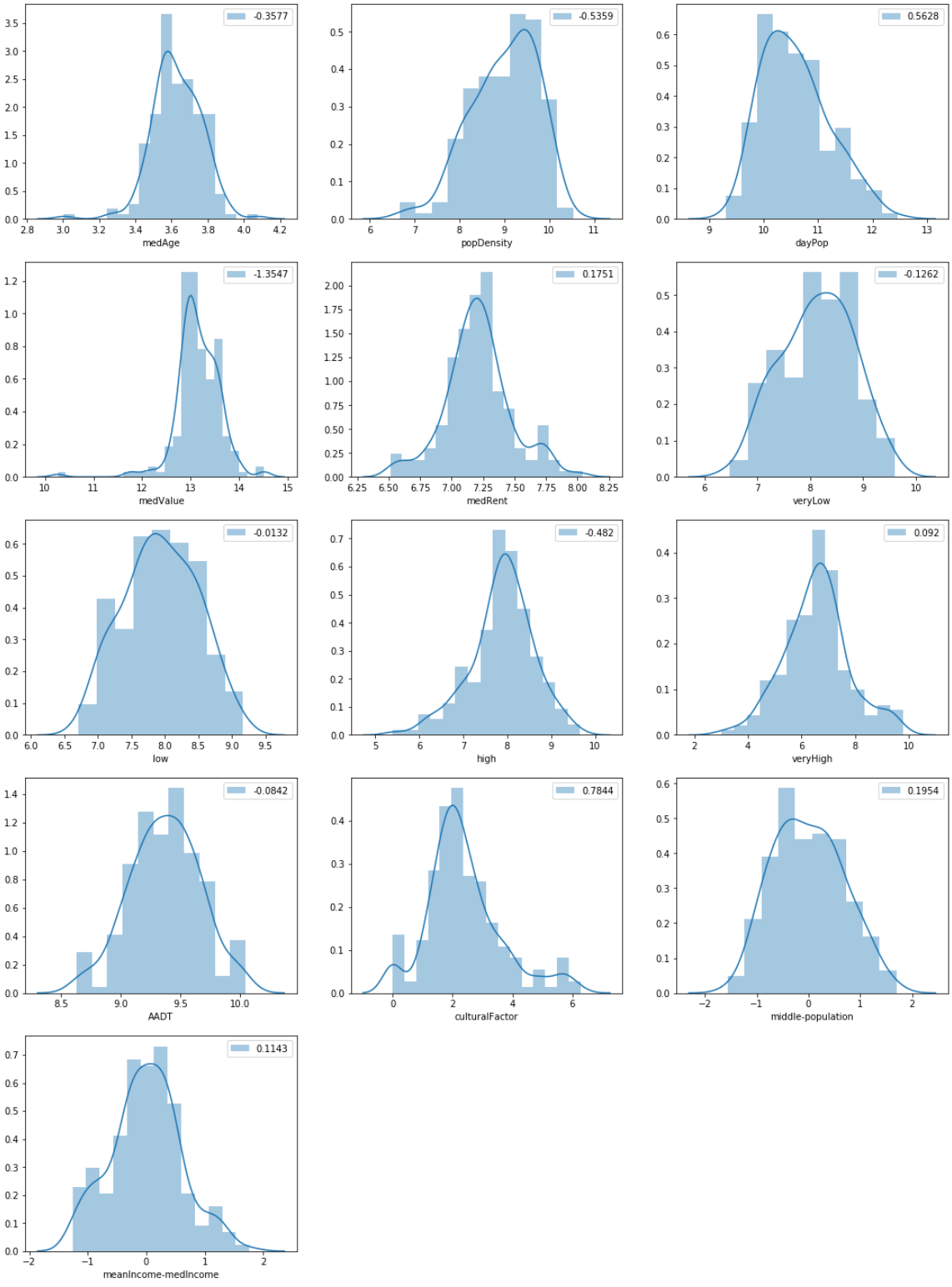
Pre-processing:

- Some of the features obtained were highly correlated with each other which led to multicollinearity. Instead of simply dropping the features, the problem was solved by using Principal Component Analysis (PCA) to reduce the correlated features into a single feature. The features with correlation higher than 0.9 were combined using PCA.
- On univariate histogram analysis, it was found that the data was right skewed. Generally, linear regression works better with normally distributed data and skewness affects results. Hence the entire dataset was logarithmically transformed and the skew was reduced greatly.
- The log Transform was applied to both features and target variables and before machine learning, large outliers were removed from the analysis by observing which entries had targets having a Z-Score of more than 3 since they might skew the analysis and affect results.



Distribution Plots of the unprocessed Feature Variables

(The distibutions are found to be right skewed)



Distribution Plots of the Processed Feature Variables:

1. Log Transformed 2. Outliers Removed 3. PCA to merge features

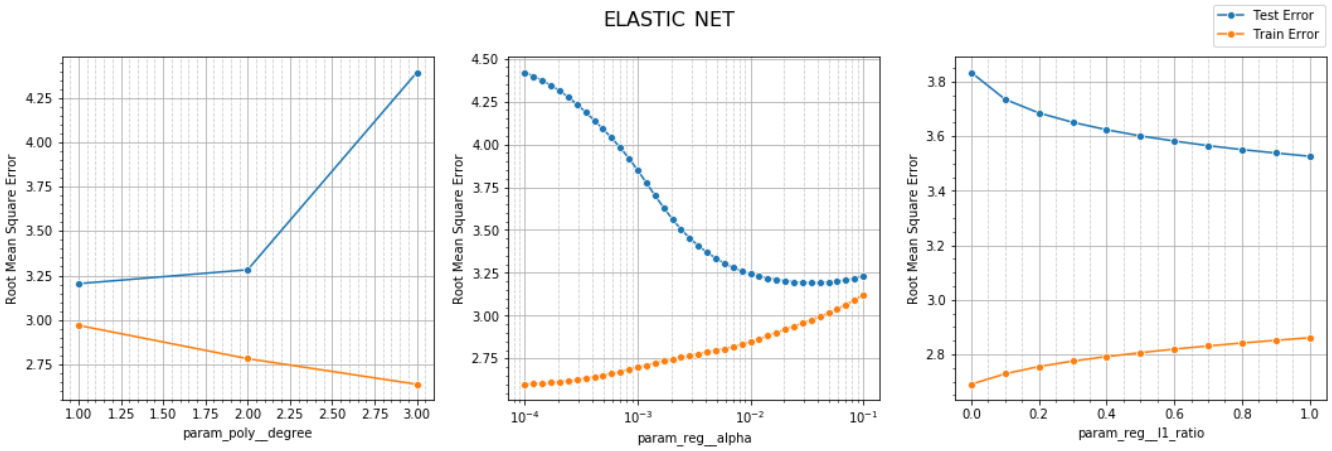
Machine Learning:

The Machine Learning process employs the following steps:

1. **Pick a Target** – The succeeding steps are performed for each target individually because each type of restaurant cannot be predicted by a single model. Separate models are used for each category of restaurant
2. **Split the Data into Train/Test Splits** – The data is split into train and test splits after the feature engineering steps are performed on them.
3. **Apply Pre-processing** – The preprocessing steps of Log Transform, PCA and removal of outliers is done on both the target and feature variables.
4. **Create Pipelines** – Pipelines are created with Polynomial Transform and Standard Scaling of the variables followed by the regressor
5. **Regression Models** – Various Regression models are developed to find the best regressor or to use a bagging method to find their combined prediction. The regression models initially developed were as follows:
    i. Linear Regression – It was ignored because of poor results for the dataset.
    ii. Elastic Net Regressor
    iii. Support Vector Regressor
    iv. Random Forest Regressor – It was ignored because of poor results for the dataset.
    v. Gradient Boosting Regressor
    vi. XGBOOST Regressor (Extreme Gradient Boosting) – A gradient boosting algorithm with regularization and other features
6. **Hyper-Parameter Optimization** – The hyperparameters of the pipelines were optimized using a combination of Randomized Search and Grid Search Cross Validation. The parameters adjusted were the degree of the polynomial transform and the parameters specific to the regressor. 10-fold Cross Validation with shuffling was done to ensure the performance of the models.
    i. Initially, Random Search was performed over a wide range of parameters to find the range to optimize in.
    ii. Then, Grid Search cross validation was performed to find the accurate parameters which provided the best results.

The best estimators using all the models were taken.

7. **Bagging the best Regressors** – The best regressors were then tested on the testing set. The predictions made by those individual regressors were given weights to ensure the best possible results and their predictions were combined to give a single prediction. The final Ensemble Regressor was used to predict the values of the number of restaurants.



Example of the curves used to optimize the regressors using Grid Search Cross Validation

Results:

Result Metrics:

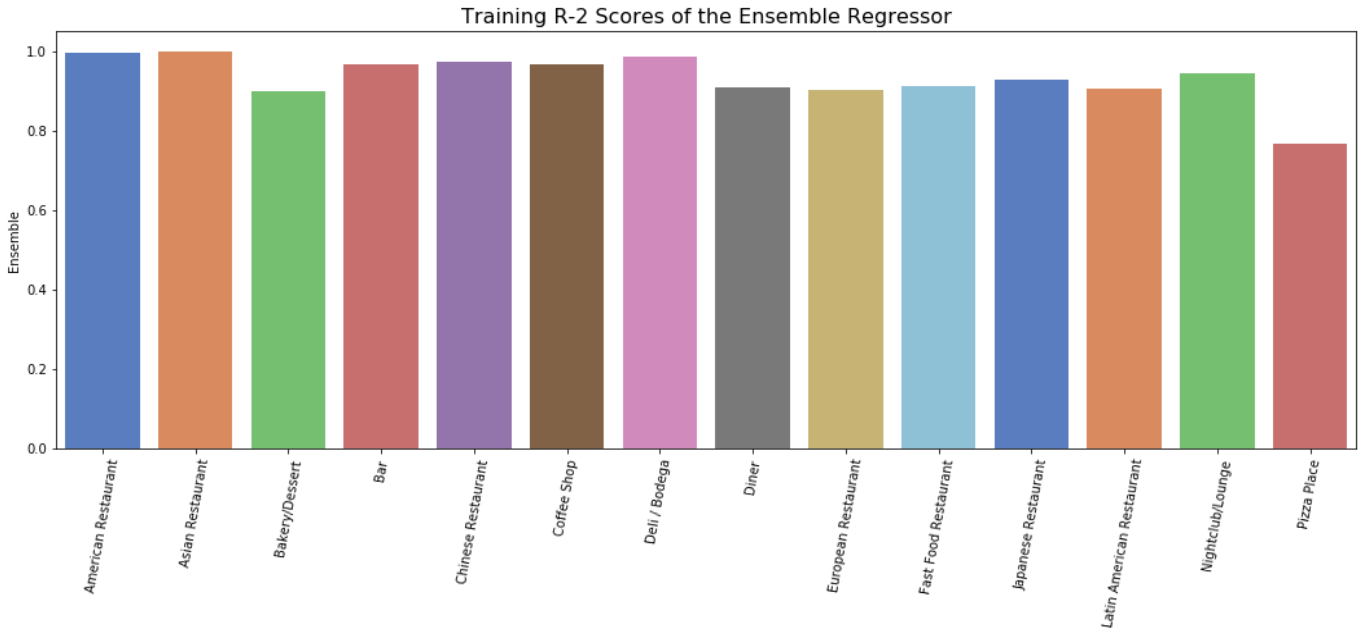
Since the targets were log Transformed before the analysis, the results are found in a different scale. In order to rectify this and to observe the results in a visually understandable scale, a custom implementation of the Root Mean Square Error was used. The Scorer first applied the inverse log transformation to each of the targets and predictions and then calculated their root mean square error as a metric of prediction. For the test set, the R2 Score, Root mean square error and the Standard Deviation of the error terms were used to assess the performance of the best estimators.

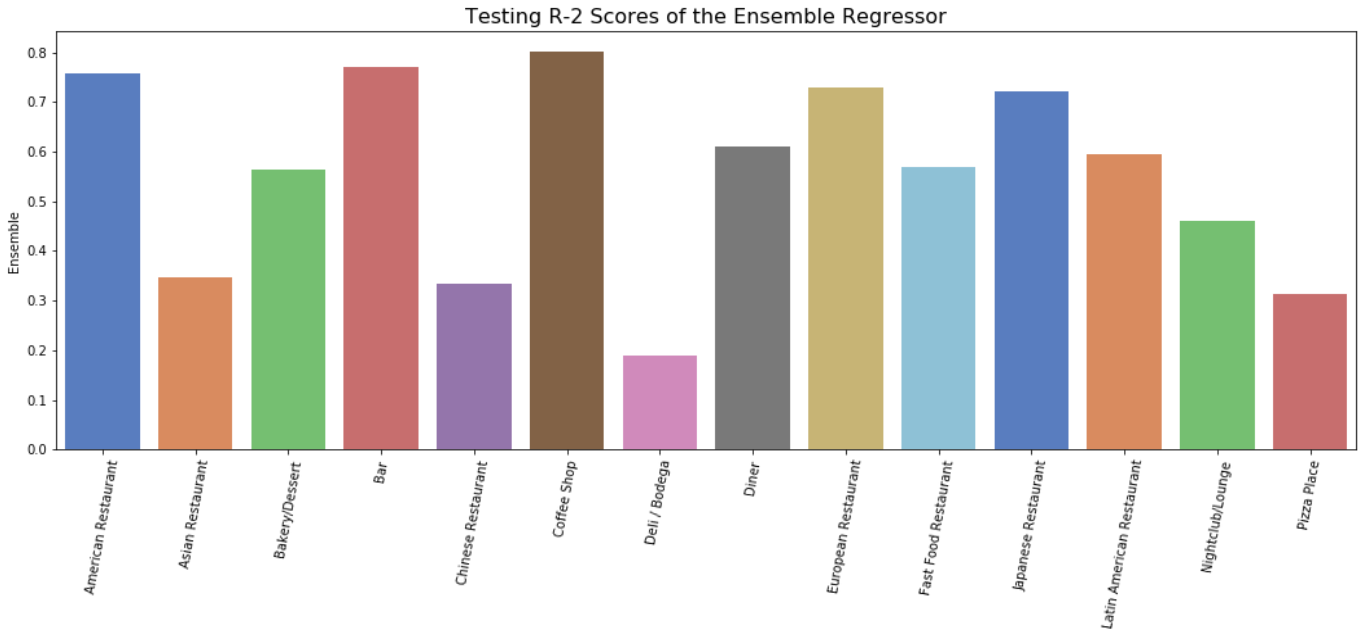
Predictor Performances:

Some of the types of restaurants had high correlation with the used features and could be predicted accurately with a high score, some of the types of targets were independent of the features and had low

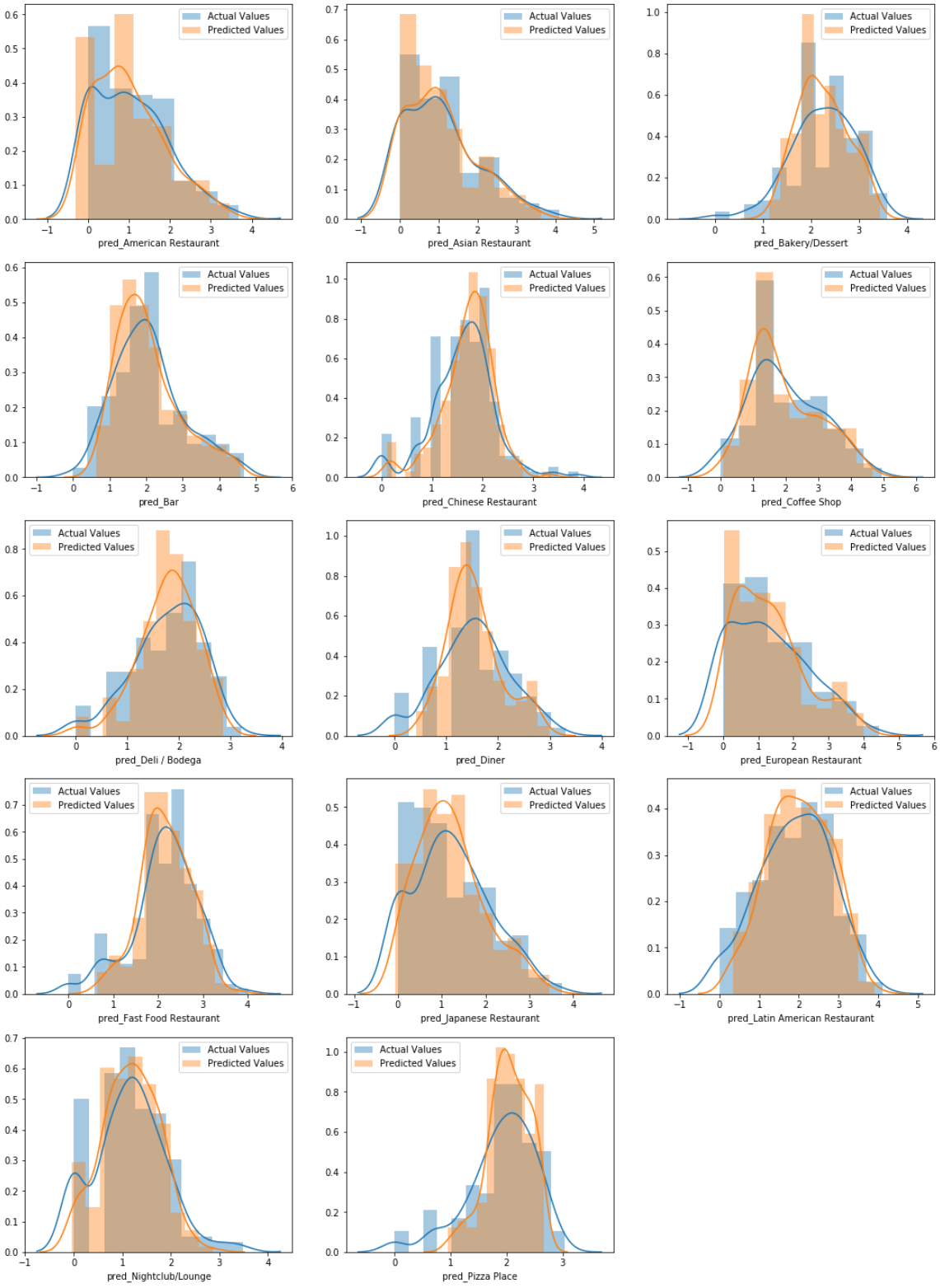
scores. Nevertheless, the model is designed to predict how many restaurants the demographic could sustain and not dependent on special features of any area.

The individual regressors have been combined to get an Ensemble regressor whose R2 Scores for the various Targets are as follows:

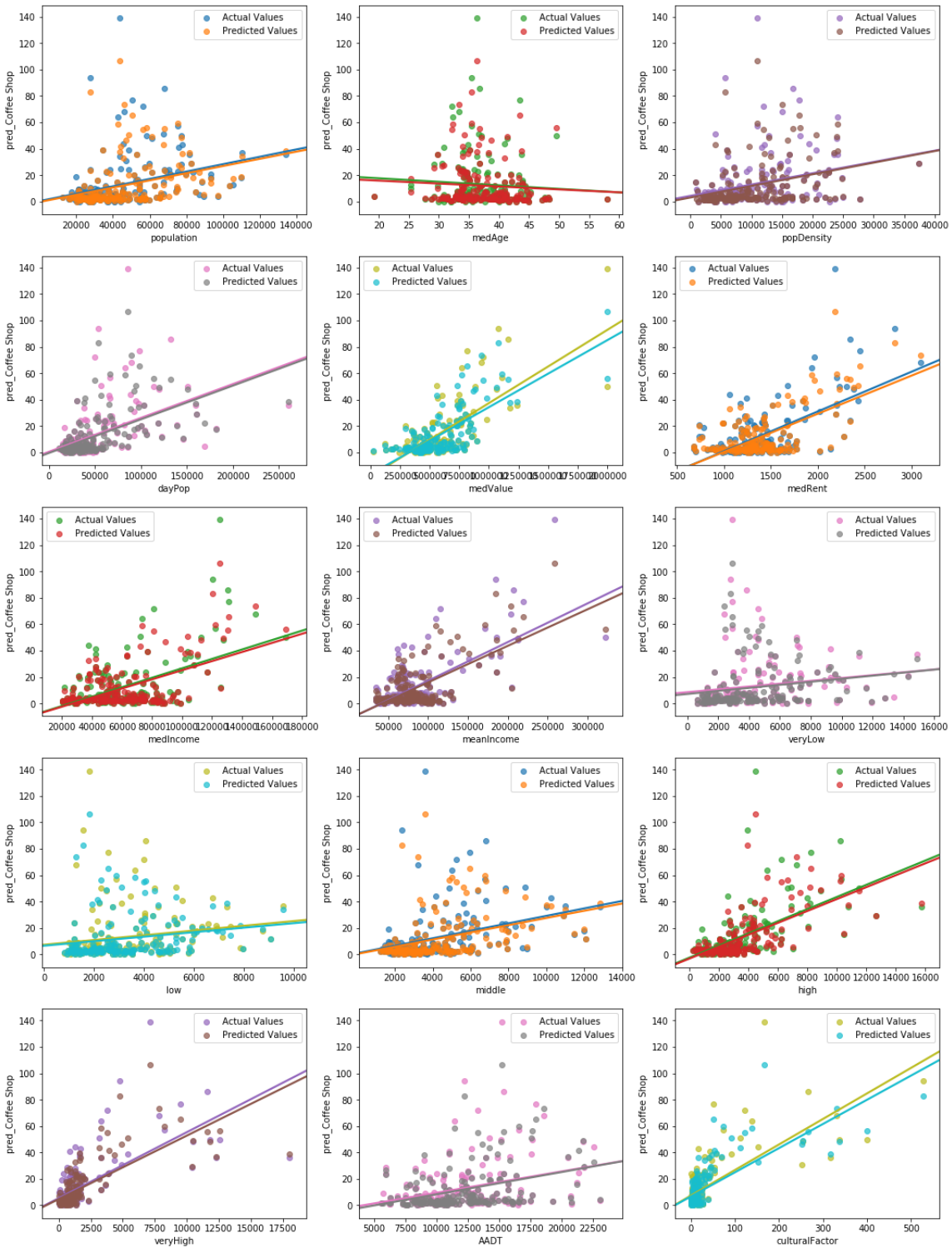




The predictions and the actual values were found similar with minor variations. The model satisfactorily indicates the number of businesses supported by a demographic. The model is then applied to actual values themselves to find where there is scope for improvement.



Log Transformed View of the Actual Vs Predicted Histograms



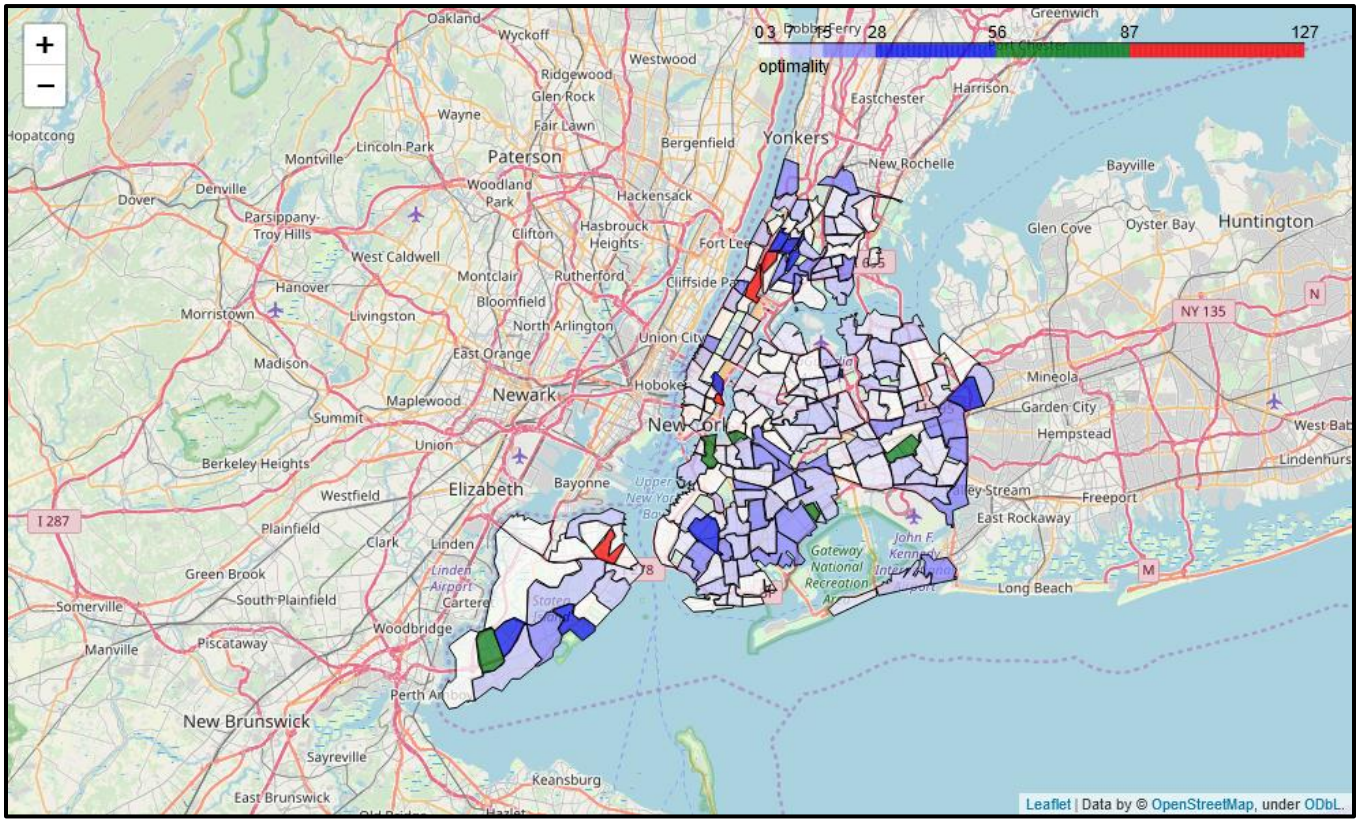
Scatter Plot of Actual And Predicted Values with each Feature Variable

(It is seen that the predicted and acutal values are very similar)

Discussions:

Using the predictions of the data by the model, we can now predict how many restaurants of a type a given demographic could sustain. In order to visualize which neighborhoods, have the most scope of improvement, we create columns representing the scope of a particular type of restaurant which is simply the difference between the predicted and actual values. *This concept should work for a new demographic where the feasibility is to be assessed and not on the training set itself*. **But citing, the lack of time, it has been illustrated on the data of New York City itself**.

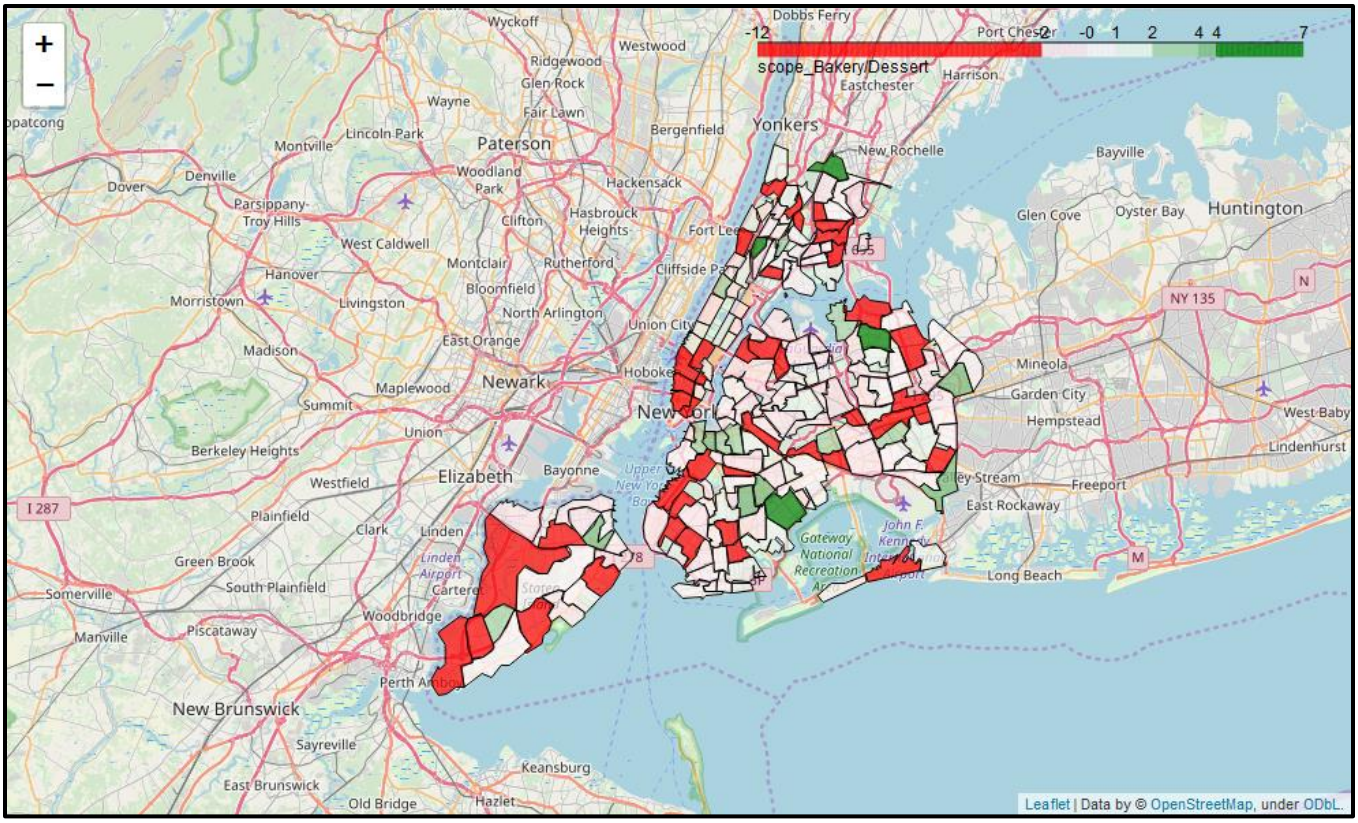
The optimality of locations has been reduced to a single variable which represents the location with the maximum scope for establishing a new business in. *The optimality is the sum of all the positive scopes divided by the total number of actual restaurants. This scaling is done to avoid locations with many restaurants taking precedence over the places with lesser number of restaurants*.



Visualization showing the most optimum locations to open a Restaurant

(Higher is Better)

The visualization may also be done to see the optimal places for a specific type of restaurant.



Visualization showing the most optimum locations to open a Bakery or a Dessert Shop

(Greens indicate optimum locations – Reds indicate bad locations – Whites indicate Neutral)

Conclusion:

This project which aims at predicting the number of restaurants in a demographic successfully does so. The entire concept of the project is solely based upon the fact that New York can be taken as a fully developed city and is at its saturation and the model would predict successfully under cases where the demographic is similar to New York. Hence it should be applicable to at least all the state capitals and big cities like Los Angeles, Chicago and such in the USA, and might even work for cities outside the US. This model should give an idea of the best neighborhoods and type of restaurant to start-up.