Likes Forecaster

Santiago Chica

Likes as a measure of the venues Popularity

One variable that FourSquare, a location data platform, has is the likes that the users gives to the venues in the platform. The likes variable can be interpreted as the popularity of each venue. Also, a venue with more likes than other can mean that the venue is more visited or the experience there was better. That´s why this project is important for the following stakeholders:

- Business owners.
- Potential Business owners.

The goal of the project is to forecast the numbers of likes of a venue, base on location, demographical, and economical data.

Data sources and wrangling

We used 4 data sources:

- **New York data:** Data of the New York boroughs and neighborhoods with their geolocation.
- **FourSquare data:** Data of the venues located in the neighborhoods of New York and their category given by the platform, likes and geolocation.
- **New York blocks data:** Data of the New York blocks. A neighborhood has many blocks, and a borough has many neighborhoods.
- **New York census data:** Demographical and economical variables of New York, given by blocks.

Final dataset

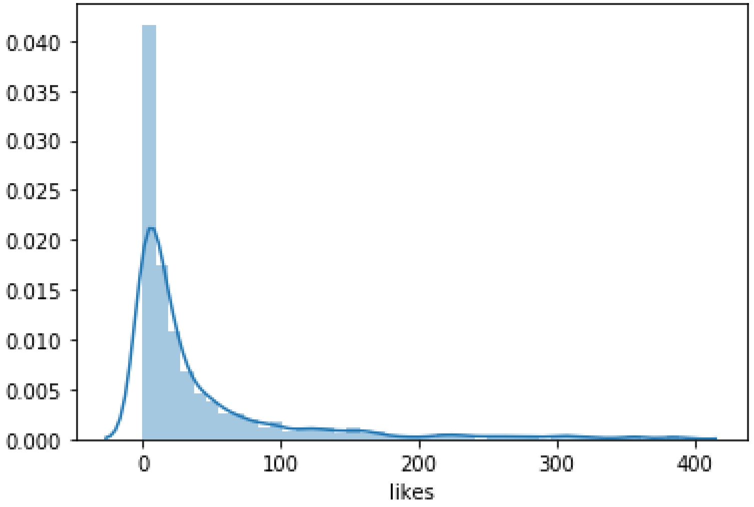
The final dataset include the following variables:

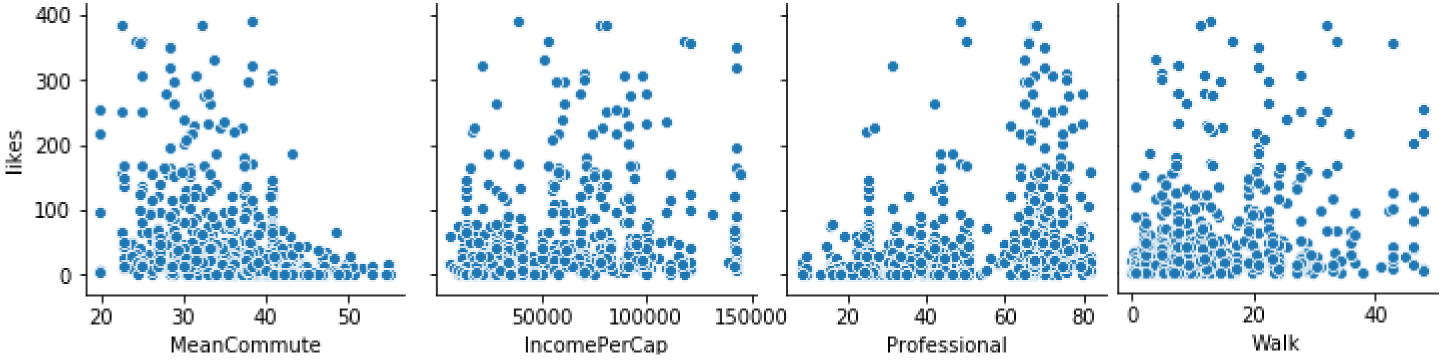
- **Borough:** Categorical variable with 5 levels. Contains the information of the borough where each venue is located.
- **Venue:** Name of the Venue.
- **VenueCategoryGroups:** Categorical variable with 6 levels. Contains information related with the category of each venue.
- **MeanCommute:** Numerical variable. Mean commute to their work of the population of the block.
- **IncomePerCap:** Numerical variable. Income per capita of the block closest to the venue.
- **Professional:** Numerical variable. Indicator of the share of the population with a professional work.
- **Walk:** Numerical variable. Indicator of the share of the population that walks to work.
- **Likes:** Numerical dependent variable. Likes per venue in the FourSquare platform.

We associated the venues information with the blocks information by assigning the nearest block to each venue. The final dataset has 753 entries.

Final dataset- EDA

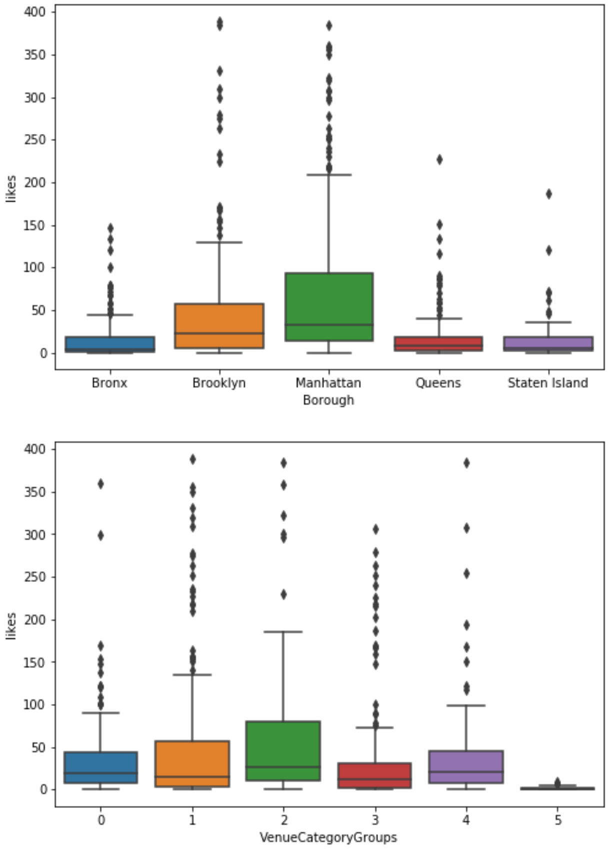
The EDA gave us some insights of the relationship of the variables. Here are the numerical variables. In the distribution plot of the likes target variable we can see that most of its values are low. The scatter plot of the likes variable against the other numerical variables show us that they don’t have a strong linear relationship.

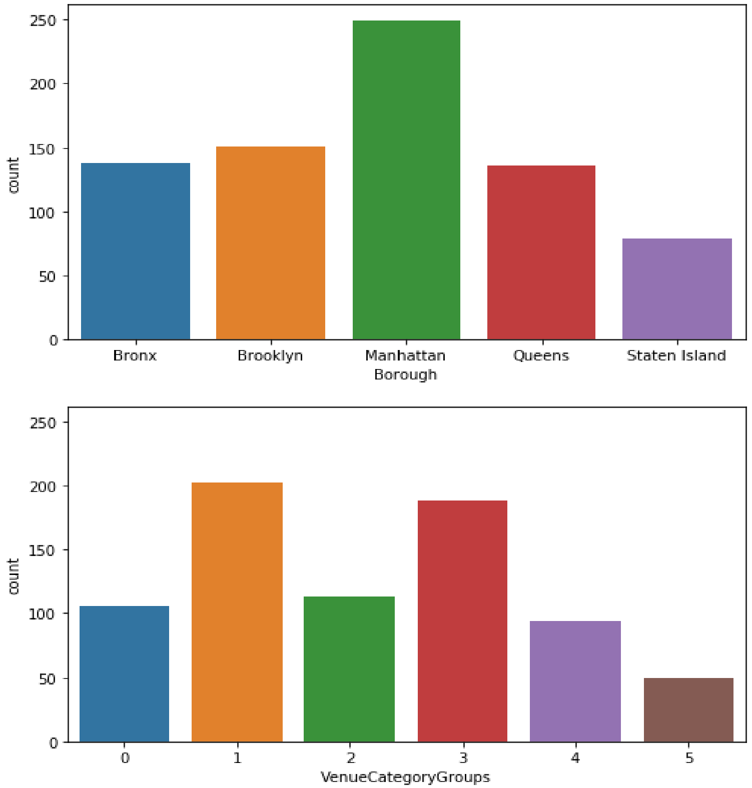




Final dataset- EDA

The categorical variables have different values in each level. Also, the distribution of likes per level is different, as shown in the boxplots. That’s why they can give valuable information to the models.


Independent categorical variables boxplot


Independent categorical variables countplot

Regression models

We used the following models for the likes forecasting:

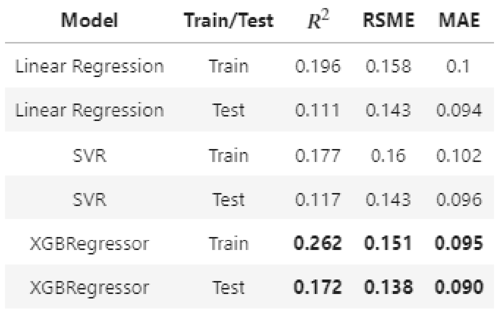
- Linear Regression. Classic linear model.
- Support Vector Regression. Classic nonlinear model.
- XGBoost Regressor. Novel nonlinear model, based on ensembles of trees models.

Their performance is measured with the following metrics.

- R squared.
- Root Mean Squared Error.
- Mean Absolute Error.

Results

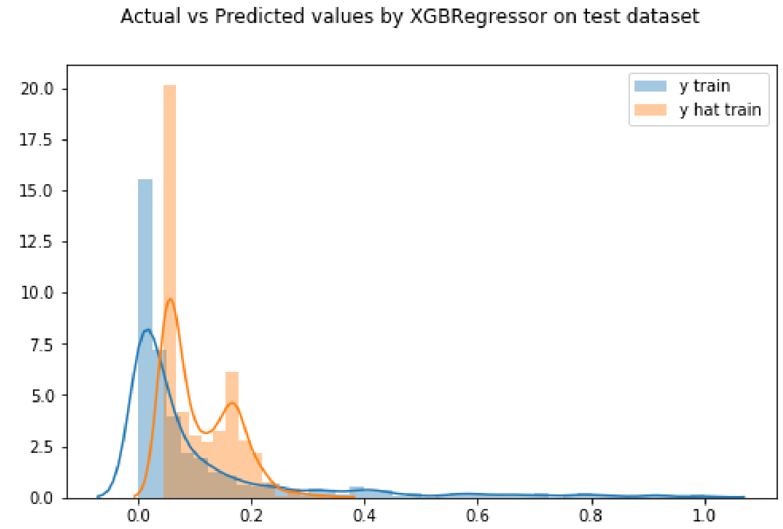
Summary of the models and their performance



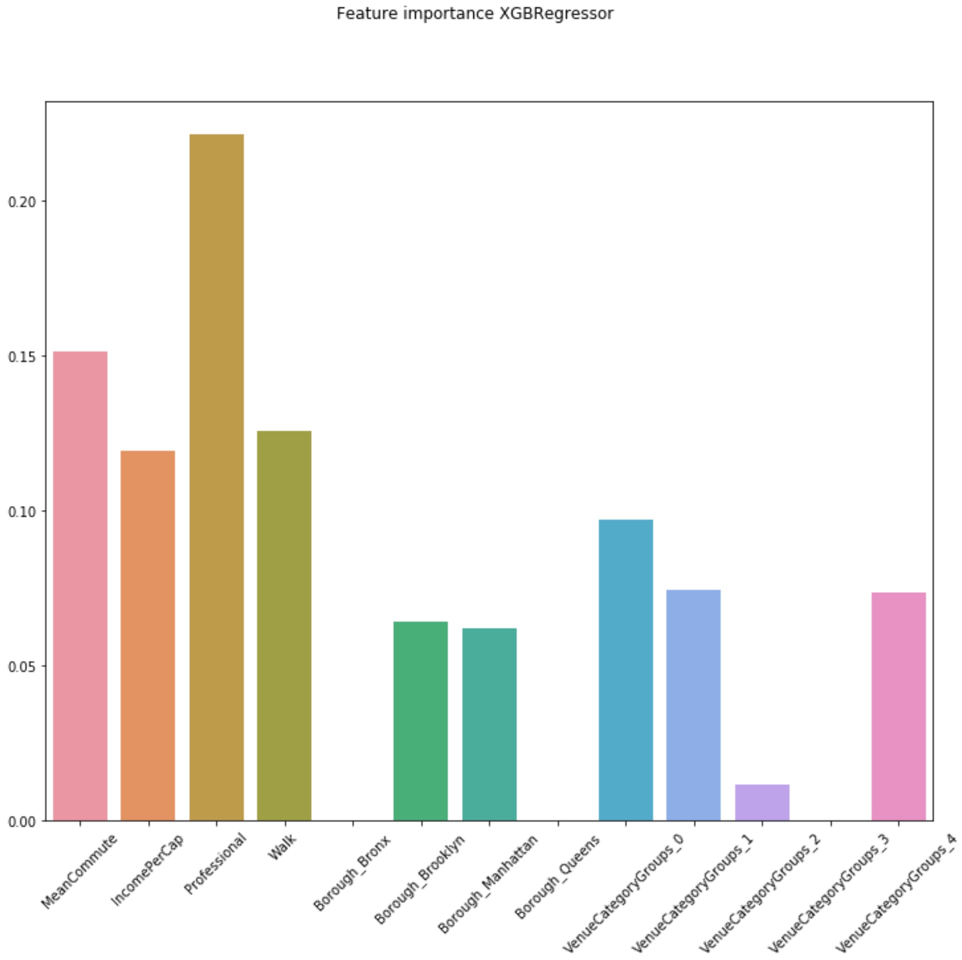
Model	Train/Test	R^2	RSME	MAE
Linear Regression	Train	0.196	0.158	0.1
Linear Regression	Test	0.111	0.143	0.094
SVR	Train	0.177	0.16	0.102
SVR	Test	0.117	0.143	0.096
XGBRegressor	Train	**0.262**	**0.151**	**0.095**
XGBRegressor	Test	**0.172**	**0.138**	**0.090**

Best model: XGBRegressor by every metric.

Distribution of the actual and predicted values by XGBRegressor


Actual vs Predicted values by XGBRegressor on test dataset

Feature importance given by the XGBRegressor. The weights range is (0,1). A higher value means that the variable is more important


Feature importance XGBRegressor

Discussion

- It seems that the regression models had a bad time trying to extract the non-linear relationship between the independent variables and the dependent variables. Since the EDA and the scatter plot we could be aware of this limitation, because regression models require a high linear relationship in order to have good results.
- The model that had a better performance extracting the mentioned non-linear relationship was the XGBRegressor, as it is an ensemble machine learning model based on trees. The metrics results confirm this hypothesis.
- Another important observation is that the models seems to overfit the train dataset and their R squared metric is worse in the test dataset. In the other metrics (RSME, MAE), the models have a good performance in the test dataset, in comparation with the train dataset.
- In general, the model with the best performance is the XGBRegressor. But if we plot the distribution of the predicted and actual values of the target variable, we see that the model is very limited. It doesn't seem that the model can output values in the whole range of the actual values of the target variable.
- Another important result is the weight of the variables in the XGBRegressor model, shown in the Results section. The variables with a higher weight in the model are Professional and MeanCommute. Some dummies categorical variables don't have a weight in the model at all. But others have a great weight. For example, the model gives a huge weight to the VenueCategoryGroups_0 dummy variable.

Conclusions and further recommendations

- The geolocation data can be combined with demographical and economical variables for constructing projects that can be solved through machine learning problems.
- Both problems of overfitness and bad performance in the metrics can be addressed by increasing the size of the dataset. In a further implementation of this Likes Forecaster, we could have a higher tier FourSquare API developer license with more calls and extract a big dataset, i.e. with more than 10,000 entries or even more. We could extract information from other cities and include them as a variable. The size of the toy set that helped us in this project is very limited.
- The most suitable model for the task of forecasting likes of the venues of FourSquare is XGBRegressor as it beats the others in every metric, in both training and test dataset. This is due to the nature of the model, as it can extract the nonlinear relationship between the variables.