Predicting Evictions

Executive Summary

Evictions are defined as a tenant, whether an individual or family, being legally, and formally removed from their place of residence by the property owner. Several factors contribute to a property owner choosing to remove a tenant from their property. hardship that evictions place on individuals and families, an analysis of evictions across states and counties will provide greater understanding of this situation.

This report looks at those factors contributing to property evictions and based on historical data surrounding evictions seeks to predict the numbers of evictions across counties. With greater insight into the number of evictions and contributing factors, actions may be taken to help mitigate evictions from happening. This report presents an overview of the underlying statistics surrounding evictions. An examination of the property values, demographics, economic factors , and ethnic considerations which contribute to eviction numbers are presented.

Data shows that certain groups, in particular lower income families, minorities, women, and those in more densely populated areas are at greater risk of eviction than others. Data was collected for two years from 1296 counties, yielding 2,546 records from which to evaluate. Analysis was done using a RandomForest Regression machine learning algorithm to help predict evictions of a test data set. Based on the findings from this analysis, and that minorities, and individuals spending a greater portion of income on housing, it becomes necessary for agencies at the state and local levels to help. Possible actions to reduce evictions may be to offer more affordable housing options, and provide services to help low income families spending a greater portion of income on housing.

Initial Data Analysis

Data was collected for two years, with 2,546 records present representing anonymous counties and states. No personal identifiable information is present in the dataset; however we do get percentages of ethnicities comprising the population. The mean number of evictions found within the data was 378, with standard deviation of 1405. The median number of evictions was 29. The data on evictions is heavily right-skewed. Evictions range from a low of 0 evictions per county, up to a maximum of 29,251. A figure f he basic descriptive statistics for evictions is presented in the below figure. The data consisted of 39 variables, plus one index column for the record identifier. Thirty-four of the features were numeric variables, all but two were continuous variables. The remaining 5 variables were categorical variables.

```
In [12]: combined[['evictions']].describe()
Out[12]:
```

	evictions
count	2546.000000
mean	378.048311
std	1405.276610
min	0.000000
25%	4.000000
50%	29.000000
75%	160.750000
max	29251.000000

Data Preparation

An initial analysis of the data indicated there were many variables with missing and incomplete data.

these features were:

pct_adult_smoking
pct_low_birthweight
pct_excessive_drinking
air_pollution_particulate_matter_value
homicides_per_100k
motor_vehicle_crash_deaths_per_100k
pop_per_dentist
pop_per_primary_care_physician

Due to the large number of missing values, it was unrealistic to fill this missing data with averages from the respective series. A decision was made to simply eliminate these features from the data.

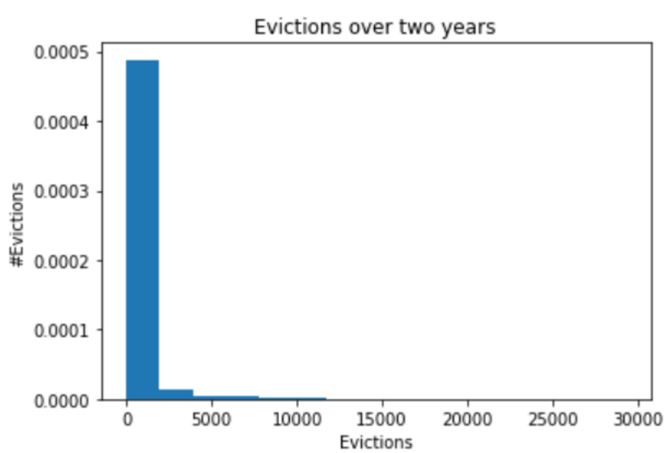
Data was also missing from the continuous variables: median_household_income and median_property_value. Of the two county records missing household income and property values, it was determined for this analysis to fill those missing variables with the mean values. There were no duplicates found within the data.

Distribution of Evictions

Initial indications show the number of evictions ranges from a minimum of 0 to a maximum of 29,251.

This is a wide range of values. We will examine the distribution of these values using a histogram

to further understand the distribution of this target variable.



The data indicates the number of evictions is skewed to the right. The average number of evictions across states from 5 to over 2000. Across counties, the average number of evictions is even larger, from

0 to over 27,000 evictions. Evidently, some counties experience significantly more evictions per capita than others. This is good information to begin examining characteristics seen among the different counties, to understand further how various demographics and economics may influence these ranges.

Population and Evictions

Another interesting finding is that on average, larger populations or more heavily populated counties tend to experience more evictions on average. This factor alone is not enough to predict evictions, as the simple fact of more people can also lead to more likelihood or chance of evictions. This will be a good feature to consider part of the analysis, but it will be used in conjunction with other influencers.

Ethnicity and Evictions

Within our sample data we are provided information relating to ethnicities of residents. Domain expertise in the matter of housing and populations often indicates that unfortunately race and discrimination often exist. We are given several ethnicities and their relative percentage of the overall population. A correlation table was checked to see whether any correlations exist between ethnicities and evictions.

Out[45]:

	evictions	population	renter_occupied_households	median_household_income	median_gross_rent	median_proper
evictions	1	0.808024	0.806802	0.13149	0.301743	
population	0.808024	1	0.974264	0.262047	0.449865	
renter_occupied_households	0.806802	0.974264	1	0.205564	0.410921	
median_household_income	0.13149	0.262047	0.205564	1	0.735655	
median_gross_rent	0.301743	0.449865	0.410921	0.735655	1	
median_property_value	0.174026	0.36291	0.360387	0.680527	0.827344	
rent_burden	0.155838	0.180068	0.168291	-0.190264	0.233079	
pct_renter_occupied	0.368408	0.353268	0.404281	-0.0501679	0.295288	
pct_white	-0.278551	-0.280231	-0.284968	0.111025	-0.212088	-0
pct_af_am	0.195535	0.105124	0.12337	-0.251263	0.0535008	-0
pct_hispanic	0.142145	0.205647	0.195405	0.045219	0.144988	0
pct_am_ind	-0.0380941	-0.0424187	-0.0393243	-0.108566	-0.0904367	-0
pct_asian	0.321428	0.510165	0.496126	0.472496	0.630138	
pct_multiple	0.0684324	0.0626958	0.0595903	0.0805218	0.156573	
poverty_rate	0.0259372	-0.0329097	-0.000166821	-0.725208	-0.371338	
pct_unemployment	-0.0200832	-0.0257856	-0.0202375	-0.487066	-0.182527	-
pct_uninsured_adults	0.0574762	0.0154079	0.0300397	-0.485359	-0.216448	-
pct_uninsured_children	-0.034523	-0.0654384	-0.0572124	-0.16543	-0.127244	-
pct_female	0.131132	0.122986	0.116302	0.0561599	0.085231	0
pct_adults_less_than_a_high_school_diploma	-0.0406337	-0.0568568	-0.0352488	-0.558119	-0.328307	-
pct_adults_with_high_school_diploma	-0.299438	-0.351491	-0.331082	-0.448223	-0.579815	-

According to the correlation analysis using scaled target values for evictions, on average it is evident there is a positive correlation between evictions and the percentage of population that is Hispanic, African American, and Asian.

Predictions

After a careful examination of our data, and cleansing and preparing, and selecting which features to include, the process of building the machine learning model begins.Data was separated into features or predictor variables, and evictions, our target variable.

Train Test Split

Scikit Learn offers a number of machine learning and data preprocessing programs. This analysis used Scikit Learn's Train Test Split program from the model selection class. The model was validated by splitting our data into a training set and test set. The training set is what the features are trained against, and a separate hold out, or test data set is created. The test size parameter was set to 30%, meaning 1782 records were used for training, and 764 held back for testing the model.

Scaling

Since our data contains values across many different units of measure and magnitude, it is necessary to scale our data, o get the values more normalized. Also,w ithin Scikit Learn is a preprocessing class called StandardScaler. This program standardizes our features using Z score scaling. This process of scaling our data helps get he data closer to a Gaussian,or normal distribution with mean of 0 and1 standard distribution. After fitting our scaler on our test data, and transforming, then the test data was transformed using our fitted scaler. The target variable for evictions was also scaled by taking the squareroot of evictions. This process helped bring the magnitude of this variable more in line with the caled training data.

Random Forest Regressor Model

As our data does not exhibit particularly linear relationships between many of our variables, a linear regression algorithm my not perform as well. A decision tree regression algorithm called a Random Forest Regressor was chosen based on its ability to use averaging to improve accuracy and control over-fitting. Additionally, this algorithm uses 'trees' to separate and make predictions based on values.

Evaluation of Performance

The R2coefficient of determination will be used to evaluate model performance. This metric is the sum of square error over the variance of our data. A value closer to 1 is est. Our analysis scored an R2 value of 0.91, indicating good predictability of future evictions based on our predictor variables. The figures below help show how the predicted target values compare to our hold out test samples.

Conclusion

This analysis shows that the number of evictions in a given state and country can be accurately predicted given the provided social economic and demographic data. The most significant features help to tell a story of what's happening. The two biggest factor are the availability of housing and the total number of people, more people leads to competitive housing which could lead to more evictions. This rising cost of gross median rent is impacting the more vulnerable in society and those had less earning power i.e. an adult with high school diploma and the elderly.