**Problem Statement**

Explore how US demographic trends influence the self-storage market. Identify the regions for future investment.

**Trend analysis**

Determine the key factor contributing to the demand for self-storage

It is difficult to quantify the demand for self-storage. Following variables have been identified as affecting the self-storage market

1. Employment

Change in the city/county's employment rate empowered many individuals to move out on their own, leading to an increase in the demand for the self-storage area.

Change in employment rate per county is considered in this exercise.

Source for the data - https://www.bls.gov/regions/southeast/data/xg-tables/ro4xg02.htm

1. Population

The need for goods and services increases as the population increases. This increases the demand for the self-storage area. In this exercise, the zip code wise population is extracted from

https://simplemaps.com/data/us-zips

1. Average household income

Rising household income affects the self-storage market; hence mean change in household income from 2010 till 2019 is considered in this exercise

Source - <https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations?select=kaggle_income.csv>

1. House Rent

The self-storage demand directly proportional to the house rent. As the house rent increases, people prefer to keep the unimportant goods into self-storage.

Source - <https://www.kaggle.com/jweinflash/us-rent-by-county>

1. Density

Along with the city's population, the city's density also plays an important role in defining the market for the self-storage industry. More the density, the more the house rent more the demand for the self-storage area.

Source - https://simplemaps.com/data/us-zips

There are other factors that affect the self-storage industry, but due to a lack of region-specific data and due to time constraints, I have decided not to include them in the exercise. These factors are as follow:

1. Nearby Universities

University students usually prefer to stay in sharing and keep extra luggage in the self-storage. Hence it is beneficiary to have self-storage area near universities

1. Nearby highways
2. Type of the people in area commercial/Residential
3. household size

Jupyter notebooks and Description

First dataset of population, density, mean house income , mean house rent and change in employment rate (county wise) is extracted and duplicated values in the records are deleted. The data files are merge on basis of zip code with the “property location” file.

1. Property\_analysis.ipynb – analysis of the self-storage property locations
2. data\_cleaning\_population.ipynb - Data cleaning for population data
3. data\_cleaning\_income.ipynb - Data cleaning for income data
4. data\_cleaning\_rent.ipynb - Data cleaning for house rent data
5. data\_cleaning\_ unemploymet.ipynb - Data cleaning for unemploymet data
6. data\_merge.ipynb – data for all the features are merge with “property location” file to prepare dataset for the experiment
7. data\_EDA.ipynb – Exploratory data analysis is performed on dataset
8. storage\_are\_prediction.ipynb – machine learning models run on the dataset and prediction is done for the city wise demand of the storage area.

Result and Analysis:

In this experiment, I used the below models to predict the demand for storage area

1. Regression model
2. Regularization models ( Lasso and Ridge)
3. Linear Regression with elasticNet
4. Random Forest
5. XGBoost
6. KNN

Model performance is evaluated on the basis of

* Average Error
* Accuracy
* Model score
* R2\_score
* Root Mean Square Error

Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Average Error | Accuracy % | Score | R2\_score | RMSE |
| Regression | 0.6489 | 94.40 | 0.63 | -1.13 | 0.92 |
| Lasso | 0.49 | 95.82 | 0.167 | 0.097 | 0.60 |
| Ridge | 0.48 | 95.82 | 0.355 | 0.09 | 0.60 |
| ElasticNet | 0.48 | 95.83 | 0.23 | 0.10 | 0.59 |
| Random Forest | 0.47 | 95.94 | 0.65 | 0.15 | 0.57 |
| XGBoost | 0.48 | 95.86 | 0.45 | 0.12 | 0.58 |
| KNN | 0.57 | 95.01 | -0.1 | -0.2 | 0.698 |

Average accuracy achieved by model is 95%

Top cities with high storage area identified with respective algorithms as follow:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithms -> | **Lasso** | **Ridge** | **ElasticNet** | **XGBoost** |
|  | Gardena | Long island | Long Island | Collin County |
|  | Long Island | Brandon | Oakland Park | Texas State |
|  | Oakland park | Oakland park | Brandon | Travis County |
|  | Charleston | Annaplis | Gardena | Harris County |
|  | annapolis | Forest Park | Annapolis | Arlington |

Future Scope:

Further, this prediction can be improved by including data related to nearby universities, population type, nearby highways.