Machine Learning Career Track Capstone Proposal:

Clustering Successful Suburbs, Predicting Home Prices and Location

Drew Lehe

Introduction:

Real Estate is a great topic for data science because

1) It's the largest industry in USA, the world's largest economy, and many other economies
2) There's a plethora of rich, freely-available data on housing
3) Built environments have a large effect on people's lifestyle, and, because housing is such a large sector of the economy, it's very influential on economies

Melbourne, Australia experienced a steady housing bubble from 2005 to 2019, fueled by a surge of foreign investment, domestic speculators, and local restrictions on land use that can cause shortages. According to the International Monetary Fund, Australia, at one point, had the 3rd-highest housing price-to-income ratio in the world, with price levels well above its historical average.

Problem Statement:

Reports are now that the bubble has burst and Australian housing prices are falling. Ideally, Australians could avoid both expensive rents and lost investment by averting a bubble altogether. Identifying what caused rising property prices will help in this endeavor.

Because of the intricate and unique nature of cities, it's essential we build a flexible model that can transfer to multiple cities and their datasets. With my skill in data mining, I hope to examine residential real estate data and find out:

-What creates a high-value suburb? Is it close to the city center and densely-populated, or is it remote and sparse? What kind of housing is built there?

I'd like to use Gaussian Mixture Modeling to create several clusters of Melbourne suburbs to find out what is driving this trend. Through analyzing the different clusters, we can find out which attributes the priciest suburbs had. With the statsmodels library in Python, I'll examine which suburbs experienced the largest growth in value. I can also use this tool to predict prices, but I'll only use it as a benchmark for comparison to the boosted decision tree regressor (XGBoost).

In addition, I'll use PCA to select the most important features, and use these features to predict which area a house is in. This tool is critical to GIS mappers who need to impute missing values, and it's crucial to marketers who need to target ad campaigns to specific areas.

Evaluation Metrics:

I'll evaluate the quality of each cluster with the silhouette score and create separate graphs of silhouette scores for each cluster, for several amounts of clusters. For the boosted tree model of

location and price prediction, I'll use classification error rate, F1-score, (classification) and RMSE (regression).

However, the important data story here is: what *kind* of housing was in each suburb? Did the high-priced suburbs have denser housing? Were there taller buildings? Or did the suburbs with pricey housing have restrictive land-use policies that only allowed low-density housing?

Project Design:

I want to create an efficient system of finding patterns in city data.

Cleaning: First, I'll load the data and clean it; outliers and unusual observations are easiest to clean with Excel, but easier to find through plotting in Python.

EDA: Using Python, I'll construct several correlograms to find any interesting patterns in the data and make note of them. Using pandas' built-in methods like .head() or .describe() helps get a preliminary feel for the data.

PCA: Next, use sklearn's PCA command to examine all numerical features, and write functions that graph feature weights and biplot all features in the 1st two principal component dimensions. Interpreting the graphs will help choose which variables to include in my clustering model.

Clustering: Now that I have clean data I understand, I'll create a *seperate* DataFrame of the suburbs, using the features I determined were important. This is where it gets interesting. First I'll cluster using KMeans, graph the results, then use Gaussian Mixture Modeling to create more detailed clusters.

Solution Statement:

By understanding what caused prices to inflate so quickly (before bursting), we can save investors and homeowners billions in losses. Experiments like this can guide municipal governments in creating policy: perhaps if denser construction ("upzoning" in planning jargon) were allowed, demand could be eased and prices would not inflate.

Marketers and other GIS mappers who rely on housing data can use this tool to impute missing data for customers and determine where to send ads.