**Housing Value Prediction in California Project Plan**

Yenchi Wang (Louis)

February 28, 2024

**Research Objective: Predict Housing Value in California**

The objective of this business analysis is to comprehend the influences on housing prices in California. We aim to ascertain the extent to which individual factors contribute to fluctuations in housing prices and explore potential correlations among these determining factors. We will be using available California housing data online from Kaggle and OpenML. By discovering these correlations, this analysis will enhance our comprehension of the investment worth of each property. Ultimately, this understanding will serve as a crucial factor in guiding our decisions for real estate investments in California.

**Business Goals:**

Our investment firm is actively seeking opportunities to invest in California's real estate market. Our strategy involves acquiring existing houses with the intention of generating profits through property rentals. Our business objectives are identifying optimal investment opportunities that can yield maximum revenue, with a focus on achieving a favorable return on investment. Our criteria for business success involves achieving a rental income that allows us to recover 10% of the investment made in our real estate properties annually.

**Data Science Goals: Measuring the correlation between each individual factor**

In this scenario, several approaches would be used to solve different types of problems we are facing in achieving our business goals. Correlation would be used to find the relationship between each factor and the housing price. Regression would be used to find the ideal properties for purchase after having the correlation data calculated. Classification would be used to identify customer/renter portrait for our properties, this would be used in the future for advertising and commercial purposes.

In technical perspective, the following ten steps would be implemented to acquire our final outputs. First step for this project would be data cleansing. The data need to be in a format that’s processable. Any missing data or typo must be fixed before moving to the next step. Second, Python would be used to read through the data set, so the data could be compared and construct into graphs and charts in the later steps. Third, the correlation between the housing price with every feature would be calculated. Fourth, comparison would be made, the factors with high correlation rate would remain while the factors with low correlation would be removed from the analysis with the housing price feature. Fifth, calculate the correlation between each factor to see if there is any strong correlation between two individual features. Sixth, if there are strong correlations between two factors beside housing price, determine the possible reason for this correlation and put it on the report. Seventh, highlight the factors that have high correlation rate with the housing price and comes up with a formula for determining the housing price with the given factors (for example, housing price = factor\_1\*(weighted rate) + factor\_2\*(weighted rate) +factor\_3\*(weighted rate) +constant). Eighth, search for the existing properties on sale and use the formula created in step seven to find properties that is worth purchase. Nineth, use classification to create different groups for the real estate properties in California. Tenth, document these three results along with charts and graphs (1. strong relations between each individual factors, 2. The formula used for housing price calculation, 3. The classification result to housing in California).

The data-science success criteria would be evaluated based on the effectiveness of the three results from the above 10 steps analysis. We need future data once our rental business starts to operate. This data will be crucial for validating the accuracy of our profit optimization formula and understanding our target customer base effectively. We will update and revise this formula frequently based on the real time data to optimize our company’s profit. The report would be addressed properly and presented periodically to the management who’s in charge of purchasing decisions on real estate.

Potential data sources would be California Housing Prices data from Kaggle (<https://www.kaggle.com/datasets/camnugent/california-housing-prices>), and California Housing Prices data from OpenML (<https://www.openml.org/search?type=data&sort=runs&id=43705&status=active>). The California Housing Prices data from Kaggle has the following features: 1) longitude, 2) latitude, 3) housing median age, 4) total rooms, 5) total bedrooms, 6) population, 7) households, 8) median income, 9) median house value and 10) ocean proximity. The California Housing Prices data from OpenML are identical to the data from Kaggle. It will be used as a back up data set for analysis. The following are two charts showcasing the features of the data that being used.

A graph with text on it

Description automatically generated with medium confidence

A screenshot of a computer

Description automatically generated

These datasets, featured in the second chapter of Aurélien Géron’s latest book ‘Hands-On Machine Learning with Scikit-Learn and TensorFlow,’ serves as resources for introducing machine learning algorithms. It enables fundamental data cleaning, features a comprehensible list of variables, and maintains an optimal size, striking a balance between being overly simplistic and overly complex. The dataset revolves around houses located in a particular California district, presenting summarized statistics based on the 1990 census data.

**Project Plan**

|  |  |  |  |
| --- | --- | --- | --- |
| **Phase** | **Time** | **Resources** | **Risks** |
| Business understanding | 1 week | All analysts and the management | Availability to all analysts |
| Data understanding | 2 weeks | All analysts | Data problems, technical issues |
| Data Preparation/ Data Cleansing | 2 weeks | Data mining consultants | Data problems,  Technical issues |
| Modeling | 3 weeks | Data mining consultants, database analysts | Data problems,  Technical issues |
| Evaluation | 2 weeks | All analysts, the management | Scheduling, ineffective model, outdated model |
| Deployment  (Repeat to Modeling phase through periodic business assessment) | 2 weeks | Data mining consultant, database analysts | Technical issues, inapplicable to real business setting |

**References**

*California Housing Prices*. OpenML. (2022, March 24).

https://www.openml.org/search?type=data&sort=runs&id=43705&status=active

Nugent, C. (2017, November 24). *California housing prices*. Kaggle.

https://www.kaggle.com/datasets/camnugent/california-housing-prices

Pace, R. Kelley, and Ronald Barry. "Sparse spatial autoregressions." Statistics & Probability

Letters 33.3 (1997): 291-297.