CPE695 Final Project Proposal

Yupeng Cao
ECE Department
Major in Applied Artificial Intelligence
10454637
ycao33@stevens.edu

Haixu Song
ECE Department
Major in Applied Artificial Intelligence
10446032
hsong13@stevens.edu

Chong Guo
MA Department
Major in Data Science
10451296
cguo10@stevens.edu

***Abstract*—This document is proposal for the final project in CPE695 Applied Machine Learning course. The project is trying to implement a housing price prediction system by using varied machine learning algorithms. The housing price prediction problem is defined in first part and we introduced the American Housing Survey (AHS) data set. Then we organized detailed implementation plan and clarify work allocation for each group member.**

***Index Terms*—Housing Price Prediction, Machine Learning, Final Project**

I. Problem statement

In this project, we propose to utilize American Housing Survey (AHS) data set to make a housing price prediction. The goal of this project is to analysis the data set and then select useful data information to build a stabilized housing price prediction system by using different supervised learning algorithm. The proposed problem belongs to logistic regression area.

II. Description of data set

A. Introduce the dataset

Microdata are files containing individual responses to survey questions. They are used to create custom tabulations, allowing users to delve further into the rich detail collected in the American Housing Survey. In the AHS microdata, the basic unit is an individual housing unit. Each record shows most of the information associated with a specific housing unit or individual, except for data items that could be used to personally identify that housing unit or individual.

B. Data Structure

There are 16 topics to be considered in our data set: Admin, Occupancy and Tenure, Structural, Equipment and Appliances, Housing Problems, Mortgage Detail, Demographics, Income, Housing Costs, Home Improvement, Neighborhood Features, Recent Movers, Delinquency, Eviction, Disaster Planning, Commuting. For each topic, there are several variables to describe them. All attributes are correlated to 'household'.

For 'mortgage.csv', it is about mortgage details. Each house has 0 or 1 mortgage. For 'project.csv', it has project details of each household and each household has 0 or more projects. For 'person.csv', it gives the people correlated to the household. Each household has 0 or more people.

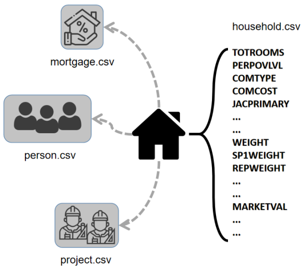

Fig. 1. Data Structure Present

III. Implementation plan

Step 1. Pre-processing: Find the outliers and missing ones, fix them or drop them due to the quality of them. Process the data according to the actual meaning of it. Generate columns of new data artificially which we think may help improve the prediction precision.Like tagging continuous values, splitting multi-tag values, etc.

Step 2. Framework Implementation: Research and determine the different machine learning techniques to build the prediction framework.

Step 3. Ensemble: Ensemble models in machine learning combine the decisions from multiple models to improve the overall performance. We would choose a voting classifier to combine the predictions coming from the models we trained.

Step 4. Result Analysing: Analyze the result in detail and visualize the important conclusions.

IV. Team members and task allocation

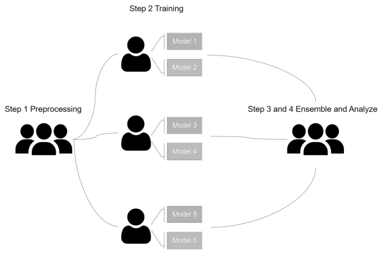

Fig. 2. Work allocation for each group member.

We train different models separately and we do other parts together.