Queens Apartment Model

Final project for Math 342W Data Science at Queens College

May 25, 2021

By Meihe Liu

1. Introduction

While New York City is among the most expensive and competitive housing markets in the nation, as one of the boroughs of New York City, the estate sale prices Queens County has a large variation every year. Zillow is one of the most popular real estate sites that estimates the sale price of houses and apartments in a given location. However, the prediction performance of zestimates for apartments in Queens is not so accurate. The goal of this paper is to build a mathematical model to predict the future sale prices of apartments in Queens.

Mathematical model is a predictive model that uses mathematical and computational methods to predict an event or outcome. Mathematical models are expressed as functions of measurable data that inherently contain error due to ignorance. To choose the "best" model, we develop a hypothesis set $\mathcal{H}$ that consists of various algorithms. This paper fits three models using Regression Tree Modeling, Linear Modeling, and Random Forest Modeling.

2. The Data

The dataset was harvested with Amazon's MTurk and raw download from their system. The raw data has 2330 rows and 55 columns, the dataset has types of data of numeric, character, and boolean. We are selecting 30 columns among 55 columns and there are 516 rows of data after removing the ones whose *sale_price* is NA.

2.2. Featurization

The variables with character type are *Title, Teward, AssignmentStatus, zip_code,* and *street_address*. The variables with factor type are *dining_room_type, fuel_type, kitchen_type,* and *season*. The variables with numeric type are *MaxAssignments, AssignmentDurationInSeconds, AutoApprovalDelayInSeconds, WorkTimeInSeconds, WorkTimeInSeconds, approx_year_built, cats_allowed, common_charges, community_district_num, coop_condo, dogs_allowed, maintenance_cost, num_bedrooms, num_floor_in_building, num_full_bathroms, num_total_rooms, parking_charges, pct_tax_deductibl, sale_price, sq_footage, total_taxes,* and *walk_score*.

3.