Team 48

David Firrinicieli, Eric Limeback, Ashwin Spencer, Andrew Taylor

Real Estate Listing Price Prediction Enhanced with Income Tax Data

Table of Contents

Problem and Approach Overview

Context & Problem Statement

CONTEXT

- Predicting house price information is not a new problem, but current approaches often don’t consider the affluency of the location of the house
- How might knowing various tax return fields by zip code, such as adjusted gross income, help predict the prices of homes for a given time period?

PROBLEM STATEMENT

We aim to improve list price prediction models that use traditional features about the home (house size, number of bedrooms, number of bathrooms, etc.) by exploring and testing the addition of income-tax related zip code features.

LITERATURE SURVEY

Existing attempts to incorporate zip code or location primarily focus on directly encoding the spatial data, but don’t account for affluency or income tax features

Research Questions & Modeling Objective

Our goal was to answer the below research questions.

Primary Research Question

How well can we predict house prices using **both** standard “listing” information about the house and income tax information about the residents of a zip code?

Secondary Research Questions

1. How does the performance of models with income tax data added compare to those with only house features?
2. Which income tax features are statistically significant? What is their relationship to house price?
3. Which traditional features are statistically significant? What is their relationship to house price? How does the set of significant features change when income tax features are added?

Methodology

The team followed the four-step approach below to develop our models and answer the key question.

Data Cleansing & Transformation

Create the dataset to feed EDA and the model

- Perform null and duplicate checks
- Confirm data types
- Aggregate data as-needed
- Join the real estate and income tax datasets and checking join veracity
- Create calculated fields for our variables as needed

Exploratory Data Analysis

Summarize and explore before modeling

- Develop univariate summaries (histograms, boxplots)
- Understand dependent-independent variable summaries (pair plots, correlation heatmaps)
- Correlation analysis

Model Development & Selection

Test and select the best model

- Split data into training/testing
- Preprocess data (scaling, PCA)
- Select and engineer features as-needed
- Train and score models from different families
- Select a 'best' model type, and fit that model on the final data

Results Evaluation

Interpret models to answer key questions

- Evaluate model performance metrics to validate results and understand the predictive power of our model
- Compare models with and without income tax data
- Develop conclusions

Process & Key Findings

Data Cleansing and Transformation

First, we completed data extraction, cleaning, and merging, including handling complexities around null values (details on following slide)

Steps Taken

- Downloaded and explored data
- Understood column meanings and selected initial feature list from income tax data
- Filtered to time window
- Removed duplicate rows from real estate data
- Corrected zip code field
- Aggregated income tax data
- Joined the real estate and income tax datasets
- Cleansed nulls and created interaction terms

Field	Description
price	Sale price of the house ($)
bed	Number of bedrooms on the listing
bath	Number of bathrooms on the listing
house	Binary flag representing whether a house or not (e.g. condo)
house.acre.lot	Lot size – evaluates to 0 for non-homes without lots
house_size	Size of the house, in sq ft
state	State
total_credit_amt	Total tax credits amount per return
taxable_income_amt	Taxable income amount per return
mortgageint_amt	Mortgage interest paid amount per return
p_mortgageint_nr	Proportion of returns with mortgage interest paid
inctax_amt	Income tax amount per return
p_unemploy_nr	Proportion of returns with unemployment
agi_amt	Adjust gross income (AGI) [2]
num_dependents	Number of dependents per return
p_re_taxes_nr	Number of returns with real estate taxes

Sources

1. USA Real Estate Dataset - https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset?resource=download
2. Individual Income Tax Statistics - https://www.kaggle.com/datasets/irs/individual-income-tax-statistics

Handling Nulls

During the cleansing process, a number of nulls were observed in the real estate data fields, and the below actions were taken.

Imputed nulls where information was available

1. For missing `**acre_lot**` data, we used the listing `**address**` to identify and assign values of 0 for apartments and condos
   - *Note: We used string matching to identify listings which included key words such as 'apt' and 'unit'*
2. In the `**bed**` field, we assigned a value of 0 to all records representing studio apartments (had 1 bath)

1,600 (21%) Records
apartments / condos with no lot and imputed value for `acre_lot`

104 (2%) Records
studio apartments with imputed value for `bed`

Removed records with remaining nulls

Field	Number of Nulls
house_size	2,202 (29%)
acre_lot	87 (1.2%)
bed	25 (0.3%)
bath	15 (0.2%)

2,329 (30%) Records
Removed

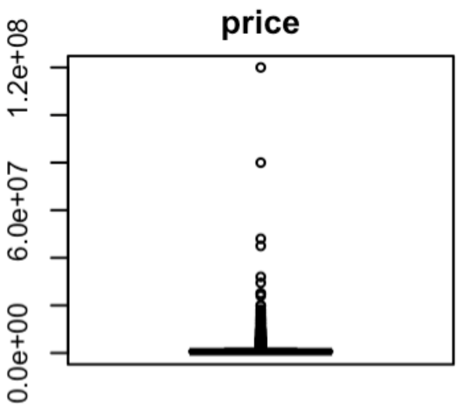
Exploratory Data Analysis

Next, we completed EDA for the dataset, evaluating individual features and the response as well as the relationships between data fields.

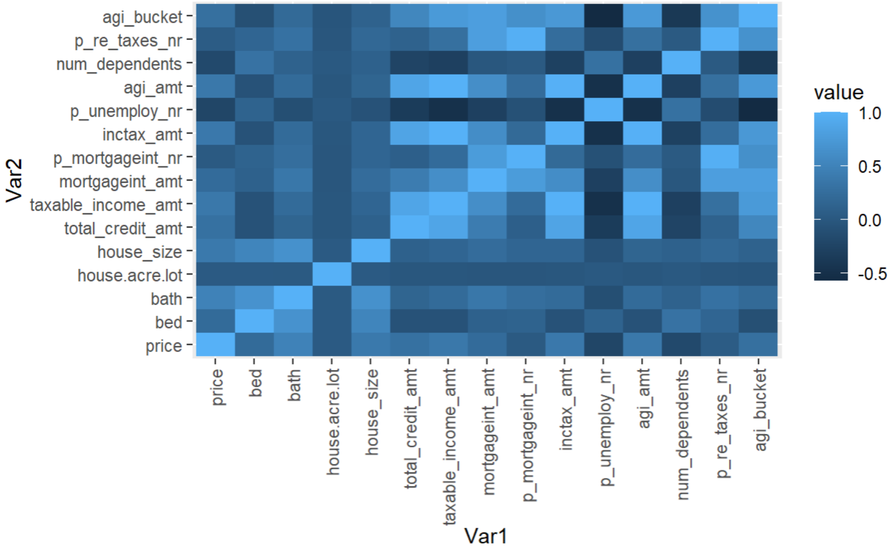
- Performed univariate analysis to explore distributions, observing and flagging any outliers for further action during modeling and analysis
- Investigated feature/DV relationships using pairplots and correlation, forming hypotheses about which features would provide the most value to the model
- Evaluated multicollinearity between predictors using a correlation heatmap, identifying whether or not mitigations measures will need to be taken

Observations and Initial Hypotheses

During EDA, a few notable observations surfaced, along with initial hypotheses.



Feature	Correlation with Price
total_credit_amt	0.290
taxable_income_amt	0.375
mortgageint_amt	0.234
p_mortgageint_nr	0.024
inctax_amt	0.378
p_unemploy_nr	-0.215
agi_amt	0.372
num_dependents	-0.179
p_re_taxes_nr	0.046
agi_bucket	0.278
bed	0.218
house_size	0.395
house_acre_lot	0.030
bath	0.490



Univariate analysis revealed outliers in the price field

Investigation of correlation between price and income tax features reveals limited linear signal

Evaluation of multicollinearity showed correlation between predictors, particularly in the income tax dataset

INITIAL HYPOTHESES

- Real estate features are expected to provide the most signal to the model
- Tax features related to affluency have the most potential to provide signal to the model
- Tree-based models will likely far outperform linear models

Model Development & Selection

During model development and selection, we outlined the model types to evaluate, prepared the data, and evaluated each model, including hyperparameter search and preprocessing as needed

Identify Model Types

- *Linear Regression* for **baseline point of comparison** and **coefficient transparency**
- *LASSO Regression* for **feature insights**
- *Random Forest* for **tree-based model benefits**
- *XGBoost* for **tree-based benefits** and **comparison to RF**

Data Prep

- Read in R packages
- Read in data and remove unnecessary columns
- Make dummy variables out of the State names
- Scale all non-binary numeric columns
- Split into training and test sets (80/20 split)

Evaluate Models

1. Perform PCA (linear regression only)
2. Build model using all features on training data
3. Build model using only real estate features on training data
4. Hyperparameter tuning for LASSO, Random Forest, and XGBoost
5. Evaluate R^2, RMSE, and MAE on test data

Model Performance Results

In evaluating the model strength, we're able to come up with a few insights around the model types and the features.

MODEL STRENGTH

	R-SQUARED	
	Real Estate Features	**All Features**
Linear Regression	0.3408	0.3938
LASSO Regression	0.3336	0.3939
Random Forest	0.3450	0.5702
XGBoost	0.2282	0.3073

INSIGHTS

- **Random Forest provides the most predictive model on both sets of features, and will be used to draw conclusions**
- For all model types, including income tax features adds between 0.05 and 0.23 to the R-squared (a significant amount!)
- Our best model has an R-Squared of 0.57, which makes it moderately strong

Feature Insights

More specifically, we can use the results of the LASSO model to see which features were selected in a regularized model.

Field	LASSO Coefficient
(Intercept)	777652.34
bed	.
bath	736419.1
house	-53034
house.acre.lot	.
state_Connecticut	.
state_Delaware	.
state_Maine	.
state_Massachusetts	.
state_New.Hampshire	.
state_New.Jersey	.
state_New.York	472845.74
state_Pennsylvania	.
state_Rhode.Island	.
state_Vermont	.
house_size	231909.44
n1_total	28145.47
total_credit_amt	.
taxable_income_amt	.
mortgageint_amt	.
p_mortgageint_nr	.
inctax_amt	334825.93
p_unemploy_nr	.
agi_amt	.
num_dependents	-226096.28
p_re_taxes_nr	-61115.92
agi_bucket	.

INSIGHTS

- The number of baths was selected, while beds was not
- A listing that's a house is ~$50k lower than one that is not, if all else is constant
- A listing that's in New York is ~$472k higher than one that is not, if all else is constant
- The income tax amount and total number of returns in a zip code both have a positive relationship with price
- The number of dependents and proportion of returns with real estate taxes both have a negative relationship with price

Conclusions

Conclusions

The models we developed help us answer our research questions.

Primary Research Question

Using a random forest model, we can achieve an R-squared value of ~0.57, which is a moderately effective model for a fairly behavioral response such as house price

Secondary Research Questions

1. Adding income tax features to the model improve the model, and increase the R-Squared value by 0.05-0.23
2. The most significant real estate features appear to be the size of the house, number of baths, and whether the house is in NY or MA
3. The most significant income tax features appear to be the number of returns, total tax credit amount, number of dependents, and the proportion of returns with real estate tax

Potential Improvements

If we were to look beyond this project, further improvements could be made to the model.

1. Enhance the data
    1. Added data to the dataset (more zip code)
    2. Add temporal features
    3. Add more detailed real estate features
2. Improve the LASSO model by adding interaction terms

Literature Citations

[1] “Machine Learning based Predicting House Prices using Regression Techniques”; J Manasa, Radha Guota, N S Narahari; https://ieeexplore.ieee.org/abstract/document/9074952

[2] “Predicting House Prices with Spatial Dependence: A Comparison of Alternative Methods”; Steven Bourassa, Eva Cantoni, & Martin Hoesli; https://www.tandfonline.com/doi/abs/10.1080/10835547.2010.12091276