MGT 6203 Group Project Proposal, Team 48

TEAM INFORMATION (1 point)

Team #: 48

Team Members:

1. **Andrew Taylor, ataylor44.** Andrew received his BS in Industrial Engineering from Georgia Tech, before becoming retail strategy consultant for the next seven years. He has previously worked on many course-related projects (this is his last semester of classes), as well as a number of primarily supply chain-related projects at work.
2. **Ashwin Spencer, aspencer6.** Ashwin has 11+ years of experience in Electrical and Systems Engineering with a focus in Computational Electromagnetics (CEM) modeling and simulation. He recently completed MS in Computer Science and is now working as a Data Scientist. He has worked on many analytics projects during his MS, and professionally, he has done probabilistic risk assessments of F-15 and F-16 engine parts.
3. **Eric Limeback, elimeback3.** Eric earned his BBA at Wilfrid Laurier University in Canada, then worked as an analyst for five years in the IT, FinTech, and SaaS industries. This is his first semester at Georgia Tech. He currently works as a Senior Data Analyst for an NLP-based chatbot company in Toronto, helping to drive strategy for some of their largest enterprise clients. Professionally, he has experience in SQL and data visualization, as well as running clustering and classification models.
4. **David Firrincieli, dfirrincieli3.** David received his B.S. from the University of North Carolina at Chapel Hill and has 3+ years of financial services experience in the Technology sector. Prior to Georgia Tech, David spent 2 years as an Investment Banking Analyst at KeyBanc Capital Markets, where he helped support and close 15+ Tech capital markets and M&A transactions totaling over $6 billion in aggregate deal value.

OBJECTIVE/PROBLEM (5 points)

Project Title: House Price Prediction Incorporating Income Tax Data

Background Information on chosen project topic:

Predicting house prices using information about the house (number of bedrooms, number of bathrooms, etc.) is a common machine learning problem. However, in many house price prediction datasets, additional information about the economic status of potential buyers is not included. When our team came across a publicly available dataset of income tax return data by zip code, we started to wonder – 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:

When homeowners list their house for sale, it can be very difficult to determine a fair valuation that might sell well in the market. No two houses are alike, and the value of the home is derived from a large combination of features like square footage, construction material, number of bedrooms, number of bathrooms, and much more. More often that not, sellers will look at comparable units in the neighborhood first to get a baseline price, which reaffirms just how difficult it can to determine a listing price based on an isolated review of the house features. We aim to improve this process by researching the extent to which additional zip code features might improve price prediction models.

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?

Supporting 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?

Business Justification:

This would be important to any company or individual who wishes to get an estimate of a home's price, without knowing what it would be listed at. This could then be used to help evaluate potential home purchases, understand the value of current assets, and inform a variety of other residential real estate investment-related activities. If I was a potential investor, I could use this to compare expected prices to actual and look for opportunities. This could also improve products which evaluate home prices, like Zestimates on Zillow.

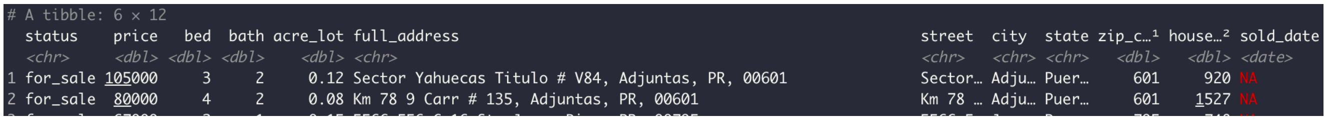
DATASET/PLAN FOR DATA (4 points)

Data Sources (links, attachments, etc.):

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

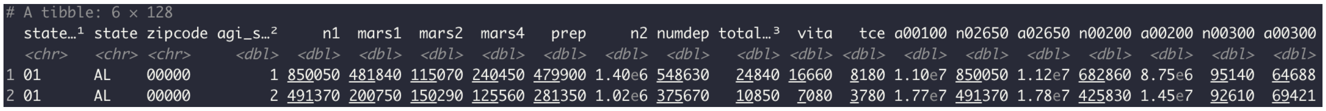
Data Description (describe each of your data sources, include screenshots of a few rows of data):

1. USA Real Estate Dataset – This is a listing and price-level dataset scraped from realtor.com, where each record represents a specific house listing (address), with multiple records if a listing's price changed. We'll use this for our dependent variable (the price) as well as the 'traditional' house listing features.



```
# A tibble: 6 × 12
  status    price   bed  bath acre_lot full_address                                          street   city  state zip_c…¹ house…² sold_date
  <chr>     <dbl> <dbl> <dbl>    <dbl> <chr>                                                 <chr>    <chr> <chr>   <dbl>   <dbl> <date>
1 for_sale 105000     3     2     0.12 Sector Yahuecas Titulo # V84, Adjuntas, PR, 00601     Sector… Adju… Puer…     601     920 NA
2 for_sale  80000     4     2     0.08 Km 78 9 Carr # 135, Adjuntas, PR, 00601               Km 78 … Adju… Puer…     601    1527 NA
```

2. Individual Income Tax Statistics – This is a zip code and AGI-bucket-level dataset generated by the Internal Revenue Service (IRS), which includes summaries of every major field on individual federal income tax returns. This includes 128 fields, which we've narrowed down to those which are relevant (listed in the next section – screenshot doesn't fit all fields). Relevant fields may then be further narrowed down using feature selection methods later in the project.



```
# A tibble: 6 × 128
  state…¹ state zipcode agi_s…²     n1  mars1  mars2  mars4   prep      n2 numdep total…³  vita   tce a00100 n02650 a02650 n00200 a00200 n00300 a00300
  <chr>   <chr> <chr>     <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>   <dbl>  <dbl>   <dbl> <dbl> <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
1 01      AL    00000         1 850050 481840 115070 240450 479900 1.40e6 548630   24840 16660  8180 1.10e7 850050 1.12e7 682860 8.75e6  95140  64688
2 01      AL    00000         2 491370 200750 150290 125560 281350 1.02e6 375670   10850  7080  3780 1.77e7 491370 1.78e7 425830 1.45e7  92610  69421
```

Key Variables:

Our dependent variable will be the 'price' field from the real estate dataset. We expect to predict the actual sale price of a home, based on the other features. Given the typical scale of home sale price, we'll likely begin without using any transformations, and evaluate the need for transformations as we go.

Our independent variables will be a combination of the provided variables in the real estate dataset, which provide listing-level data, and others from the income tax dataset, which provide zip code-level data. Since the income tax dataset is initially at the zip code and AGI-bucket level, we will need to perform aggregations to get to our features. These aggregations generally will calculate either an amount per return or a proportion of total returns in a zip code. Since the real estate dataset may have multiple records for a home, we expect to remove duplicates then take the latest price for a home. Below are the initial features we expect to derive from our datasets and their expected relationship to the house price.

Although we don't currently plan to create any interaction variables or transformed variables, we may decide to incorporate these once we begin to analyze the data. For example, we may theorize that the number of bedrooms has a different relationship with price in zip codes with more dependents on their returns (larger families), and so we may introduce an interaction term of (beds * num_dependents) to evaluate that hypothesis.

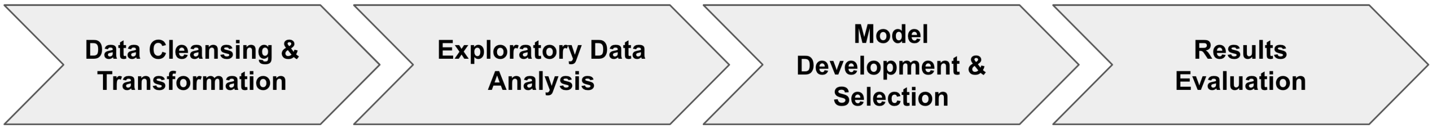
From the real estate dataset, we would expect that the number of beds, number of baths, and house size would all be significant. From the income tax dataset, there's a lot to evaluate. Although more analysis is needed to make a definitive statement, we expect the most significant features to be those related to affluency – taxable income, AGI, total income tax, and total income amount.

	Variable	Description	Expected Relationship to DV
Real Estate Data (Listing-Level)	**bed**	Number of bedrooms on the listing	The more bedrooms, the higher the price
	bath	Number of bathrooms on the listing	The more bedrooms, the higher the price
	acre_lot	Size of the lot, in acres	The large the lot size, the higher the price
	house_size	Size of the house, in sq ft	The larger the house size, the higher the price
	state	State	Nonlinear - encoding the state would allow the model to adjust prices at a state-level, if all else is constant
Income Tax Data (Zip code-Level)	**total_credit_amt**	Total tax credits amount per return	The more tax credits, the more affluent - the higher the zip code's prices are
	taxable_income_amt	Taxable income amount per return	The more taxable income per return, the more affluent - the higher the zip code's prices are
	mortgageint_amt	Mortgage interest paid amount per return	The more mortgage interest paid per return, the more house people are buying - the higher the zip code's prices are
	mortgageint_nr	Proportion of returns with mortgage interest paid	The higher proportion of returns with mortgage interest paid, the more homeowners - the higher the zip code's prices are
	netinvest_inc_amt	Net investment income tax amount per return	The more net investment per return, the more capital in people's pockets - the higher the zip code's prices are
	netinvest_inc_nr	Proportion of returns with investment income tax	The higher proportion of returns with mortgage interest paid, the more people with capital - the higher the zip code's prices are
	inctax_amt	Income tax amount per return	The more income tax per return, the more affluent - the higher the zip code's prices are
	unemploy_nr	Proportion of returns with unemployment reported	The higher the proportion of returns with unemployment, the less affluent - the lower the zip code's prices are
	total_income_amt	Total income per return	The more total income per return, the more affluent - the higher the zip code's prices are
	agi_amt	Adjust gross income (AGI) [2]	The more AGI per return, the more affluent - the higher the zip code's prices are
	num_dependents	Number of dependents per return	The more dependents per return, the more families - the more house size and bed/beth may matter
	re_taxes_nr	Number of returns with real estate taxes	The higher proportion of returns with real estate taces, the more people with capital for real estate - the higher the zip code's prices are

APPROACH/METHODOLOGY (8 points)

Planned Approach

We plan to break this problem into four main steps, which may be iterative depending on our findings.



Data Cleansing & Transformation

Exploratory Data Analysis

Model Development & Selection

Results Evaluation

In the data cleansing and transformation step, our objective is to create the dataset which we'll study in exploratory data analysis and ultimately feed to the model. The activities in this step include performing null and duplicate checks, confirming data types, aggregating the daily datasets to an annual level, joining the datasets and checking join veracity, and creating calculated fields for our variables as needed. Specifically, we know that we need to transform the income tax data by aggregating it to zip code-level and calculating the ‘per-return’ and ‘proportion of return’ fields in the table above. We also know that we’ll need to remove duplicates and take the latest listing from the real estate dataset.

Next, in the exploratory data analysis step, our objective is to summarize the data in a way that will help us understand anything we need to know before modeling. The activities in this step include univariate summaries (histograms, boxplots), dependent-independent variable summaries (pair plots, correlation heatmaps), and any other simple summaries to get to know the data better. During this step, we may find outliers or other phenomenon in the data which require additional transformation.

In the model development and selection step, our objective is to test various model types and select the model which provides the best results. We will likely test model types whose coefficients can be interpreted (linear regression, LASSO regression) as well as opaque models (Random Forest, Boosting). Activities here include splitting data into training/testing, scaling data as needed, performing PCA if needed for dimensionality reduction, feature selection, training and scoring models from different families, selecting a ‘best’ model type, and fitting that model on the final data. Specifically, we’ll likely pick regression-appropriate error metrics for comparison (mean squared error, R-squared), and we’ll use grid search and cross-validation to train and tune the hyperparameters for our models.

Finally, in the results evaluation step, our goal is to interpret the model to answer our research questions. The main activity will be evaluating model performance metrics to validate results and understand the predictive power of our model. In addition, it’s during this step that we’ll run a version of the model without the income tax features to understand the value that those features add. We expect that all analysis will be done in R Markdown files.

Anticipated Conclusions/Hypothesis

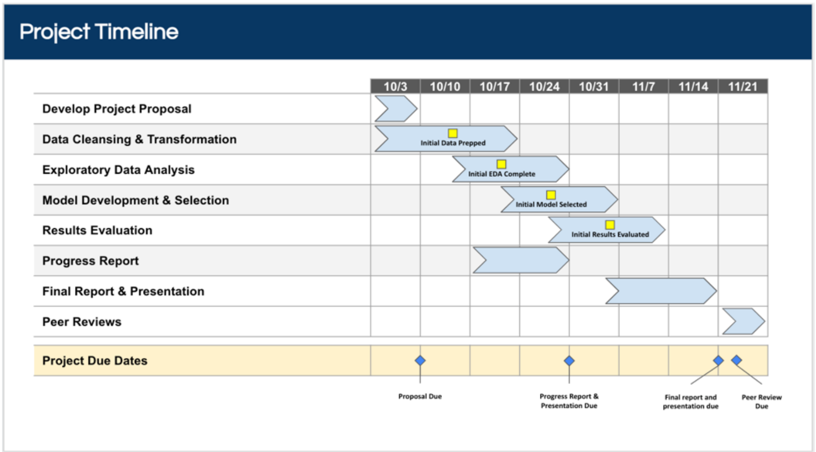
In order to answer our primary question, we’ll review the error metrics to understand how well the model performs. To understand how much value income tax features add, we’ll compare the final model to a version without the features. Finally, to understand the relationship between features and the price, we’ll review our best “transparent” model (e.g., multiple linear regression) to evaluate the coefficients.

What business decisions will be impacted by the results of your analysis? What could be some benefits?

As mentioned above, this would provide additional predictive power to any business related to real estate investing. In addition, this would hopefully allow models to be more localized, giving better predictions to specific zip codes.

PROJECT TIMELINE/PLANNING (2 points)

Project Timeline/Mention key dates you hope to achieve certain milestones by:



In order to meet the project deadlines stated in the Project Instructions, we’d follow the calendar to the left. This would begin with completing the first round of data cleansing & transformation by 10/13, in order to begin EDA. Similarly, we would complete initial EDA by 10/20 in order to begin model development, select an initial model by 10/27 in order to begin results evaluation, and have all modeling complete by 11/10 to complete our final report. This approach frontloads the work, in case the inevitable roadblocks pop up. It also allows for a full 10 days at the end of the project to complete the final report and presentation.