**Real Estate Price Predictor**

**Presentation Talking Points**

**Slide 1**

Real Estate Price Predictor

Group: Arlie, Daniel, Johnny and Baris

With economic and financial markets changing rapidly, there is a high demand for understanding current and future real estate price predictors. Since Covid, the real estate market rapidly excelled due to high homeownership

Real Estate Price Predictors such as the Zillow Zestimate was embraced by consumers, because it has allowed Sellers to gain valuable information about their home's worth with a single click.

Also, it has assisted Buyers to become more savvy in knowing a property’s worth before putting in an active offer for purchase.

Before, this process may have involved identifying a lot of individual data points and talking to a local Realtor.

The current main price predictors are using close to real time real estate data and adding other current economic indicators in order to predict price.

**Slide 2**

Executive Summary

Our focus was on Real Estate

Benefits of using Tech:

* **Reduces Risks:** estimate the overall condition like its age, deconstruction history, owner information. the company can provide their customer with up-to-date information so it increases their satisfaction from working with them.
* **Data-driven decision:**Machine learning open many opportunities for the business. just feed the algorithm with data and it will process it to help you make the right decision.
* **Marketing strategy:** with the help of customer information company can plan their future marketing strategy according to customer needs.

**It helps calculate the price:** *current and forecast the future of the market*

Goal was to use data science and machine learning to predict current real estate prices.

* Can we predict housing prices based on this data?

**Slide 3**

We want to predict price using variables such as Bed, Bath, House Size, Lot Size

Selected a Linear Regression Model as we need to predict a continuous dependent variable using a set of independent variables

This means that we have a regression problem that can be solved using linear regression model

**Slide 4**

We imported a “USA Real Estate Dataset” from Kaggle

This dataset contains Real Estate listings in the US broken by State and zip code.

Data was collected via web scraping from [Realtor.com](http://Realtor.com) using python libraries.

This data is updated weekly.

The dataset has 1 CSV file with 12 columns -

realtor-data.csv (400k+ entries)

* + status
  + price
  + bed
  + bath
  + acre\_lot
  + house\_size
  + full\_address
  + street
  + city
  + state
  + zip\_code
  + sold\_date

Collection - imported data and turned into Dataframe

Cleanup -

Inputs had multiple categories (as stated above)

Dropped all categorical data

Especially zip code

Took all fields and created new state dataframe

0 if False, 1 if True

To avoid colinearlity issues

Reduced the number of inputs by dropping 2 states Virgin Island, Puerto Rico

Data fit the model a bit more

Concatenated two data frames together to be the training data for our ML model

Dropped the null values

With this certain states needed to be dropped as they ended up having no data

Georgia, Virginia and a couple other states

Preparation & Training Process-

Price is our target (output)

Define Y = price as the target

X is all the training data and drop target outputs

Train, Test, Split

Linear regression model to instantiate the model

RECHECK NUMBERS!!!!!

Training data 16% accurate

Drop bed- test data smaller difference in accuracy

Drop bath- was larger difference in accuracy- VIF score was a 17

Anything above 5 take a look at

Heat Map

Strong Association between neighboring states = Rhode Island and Mass, Vermont & Mass

Weak Association between NY and acre lot

Weak Association Beds and House Size

**Slide 5**

Random Forest-

Mean Absolute-

Mean Squared -

Large Mean Squared means there is a lot of noise

**Slide 6**

**Slide 7**

Real Estate price predictors are only as accurate as the data behind them, meaning they may be outdated or incorrect.

Our dataset had only minimal real estate property factors (Bed, Bath, House Size, Acre Lot, State, Sold Price), while multiple other data sets include factors such as longitude, latitude, pool(size), garage(size), ocean/body of water proximity, major highway proximity, median house value, median income, view, overall condition, total floors, etc.)

Best idea would be to make our own, constantly updated (weekly) data set with many more factors/inputs. What was easily available and free was not substantial.

* Api integration - increase the accuracy of our predictions with real time data
* Build out a better user interface to help anyone who is looking to research prices and get an understanding of where the market is headed
* Implement tools (specifically  to help our model take into consideration factors like gdp, inflation and economic cycles to make our predictions even more accurate