Property Sales Script

- Slide 1
- My name is Moshe Burnstein and I will present to you my project called Predicting Property Values in New York City. I aim to build a most robust model which captures the important features to confidently predict property value.
- Slide 2
- I will state the Business Problem, give some background, present some models, give an analysis of the modeling, discuss the ethical implications, and list my references.
- Slide 3

- The business problem is how can we better inform buyers and sellers of property throughout the five boroughs of what the true value of that property is? This information is invaluable to all the principals involved in real estate transactions. The real estate market at large stands to gain greatly from good predictions because as the saying goes ‘clarity builds trust’. This in turn will make the market all the more robust.
- Slide 4
- There are many acknowledged drivers of property value. These include location, location, location, square footage, comparable property values

as assessed by their sales, condition and age of the property, and interest rates. Another driver is known as the income approach which values a property by how much money it will bring in. The problem to solve is what mix of these factors best predicts the actual value of the property.

- Slide 5
- The data I will use comes from the City of New York. It includes all property sales from September 2016 to September 2017. There are 84,548 entries, but most of the rows are unusable. There are many nulls and 0 values for Sale Price. The 0 Sale Price indicates a sale without

consideration, which means that something of monetary value is given for something which does not have monetary value. A common example of this is a parent gifting a property to his child for 0 dollars. This type of data is useless for us because it does not tell a story of property value. I excluded all Sale Prices less than $5,000 because they are unrealistic values. After all the wrangling I was left with a bit more than 29,000 sales records. These entries include number of commercial and residential units, square footage…both land and gross, and borough. I used dummy variables to

one-hot encode the categorical variable borough.

- Slide 6
- This correlation heatmap shows poor correlations between the independent variables and the target…Sale Price, except for gross square feet which is positively correlated with sale Price. One must be wary of the correlations between features such as residential and commercial units and total units, and between land square feet and gross square feet.
- Slide 7
- Manhattan, with an average Sale Price of almost $20 million, far

outclasses all of the other boroughs. Its average number of commercial units also outpaces the other boroughs by orders of magnitude. One may choose to look at property sales in Manhattan separate from the other boroughs.

- Slide 8
- Here are 2 of the many models that I built. The classical linear regression did not fare poorly, but the Random Forest is my most robust model. Across the board, it outclasses the linear regression. The R-Squared of the RandomForest is almost 59% to the 49% of the Linear Regression. In all of the error metrics the

Randomforest scored lower. RandomForest is an ensemble model which merges the predictions of multiple Decision Trees.

- Slide 9
- I used my model to make predictions for each borough and I gained insight. I set each of my features to its mean and added one or subtracted one from one feature. I had presumed that adding one commercial or residential unit to the mean while keeping all other features at their mean would add value. And subtracting one commercial or residential unit from the mean while holding all other features to their

mean would lessen value. I found this not to be true for Manhattan, Staten Island, and to a lesser extent The Bronx. Lesser units for the same square footage commanded higher prices. This may impugn the ‘income approach’ mentioned earlier. This most definitely demands more investigation.

- Slide 10
- I enthusiastically encourage you to make confident predictions with the RandomForest model. All you need to do is feed the model with your specifications of features and it will output a value that you can leverage knowing how robust a prediction it is.

You must constantly monitor the model to keep it up to date. And always explore new potential features. The real power of this type of modeling is to supplement your expert knowledge and expertise in the real estate market, the effects of interest rates, and the current political climate and put all this knowledge together to make an accurate prediction.

- Slide 11
- Ethical concerns must be dealt with. There must be clear documentation detailing the rationale for wrangling the data. Every omitted or imputed row must be justified. One must

ensure that no personal data is used unlawfully. One must clearly explain the models and what the true import of a confident prediction means.

- One must be cognizant of the real estate market's poor history of discriminatory practices... from 'redlining' black households as too risky to lend to, to predatory lending. One must beware of causing gentrification, even if it will potentially cause us to make a lesser profit. One must engage the local community and keep open these lines of communication.
- One must periodically perform ethics audits where the actual company

activity is compared with their written code of ethics.

- Slide 12
- References
- Slide 13
- Thank you for attending my presentation and feel free to reach out to me with any questions or comments.