Slide 1

Hi all! My name is Jesse and welcome to my presentation! Thank you very much for being here.

Just a brief introduction about myself; I came from a music background. I play the double bass and was the youngest member of the Vancouver Symphony Orchestra when I joined them in the capacity of a substitute player. I've travelled all over the world for concerts and festivals; however, I am getting married this September and with the future prospect of starting a family with my fiance, I've decided to pivot into the field of data science; something that I've always been interested in and exploring, that will also offer me the stability in employment that i am looking for.

As you can see here, my presentation will be composed of three segments: I will first define the business case that I have been tasked with by Bridgestone Realty, the real estate company i've been working with for this project. Second, I will go over some of the data science aspects of the project, and finally, I will do a brief demo of the application that I have been developing for my Vancouver real estate property estimator.

As I mentioned just now, I have been very fortunate to work with Bridgestone realty, a real estate company that specializes in foreign investors and property management in Vancouver. They granted me access to the Paragon MLS database, which is the proprietary source of information for realtors in Vancouver, to help them create a dynamic landing page application that can produce accurate estimates of any property within the Greater Vancouver Area.

We have identified two major advantages of creating this platform.
First of all, an instant, convenient and always available estimator can serve as a great introduction for anyone who is interested in making their first step in the realty market. The traditional approach for anyone thinking about realty involves clearly defining their want and need, doing research and comparing existing listings of property that share similar characteristics, and then setting up appointments with different realtors before they can make any decisions. The application, with its ability for the consumer to input any particular characteristics of the property they would like, does away with all the complexities mentioned above. The consumer then, is much more likely to take the first step and proceed in their collaboration with Bridgestone Realty.

Secondly, the question of conflict of interest is a big concern for many buyers and sellers when they work with realtors; I can attest to this myself and in my experience of talking to various people, both customers and realtors, on the subject. We are always worried that the realtor will set the price too low to have a bigger chance of gaining their commission when selling a property, or doing some other things for their own business interest. A mathematical machine learning model can assure the customer of this worry, since it's not biased by any personal agenda from the realtor in its estimation.This can also be a display and commitment to transparency and clarity from the realty company if it is featured on their landing page.

Now I shall move on to the second segment of the presentation, which will be a brief overview of some of the technical, data science aspects of the project.

I am very glad to work with Paragon MLS, which provides very high quality data for Vancouver realtors. I set my scope on all real estate transactions within the past year in Vancouver, which comes to around 65000 observations. They include very detailed information such as the history of each property, their days on market and activities, their proximity to public transportation and schools, and their zoning status and legal standing with the various city governments in the Greater Vancouver Area. I used feature engineering techniques such as embedding, tokenization and sentiment analysis to extract suitable information, which comes to 24 features in the end. After this initial preprocessing, the data then goes through a pipeline, where they are normalized with standard scaling. These functionalities fall under the framework of sci kit learn. Finally comes the fun part, which is building neural networks and using machine learning algorithms, and pitting them against each other to see which performs the best.

As you can see on the graphs here, these are some of the machine learning algorithms performances, ranked by their mean average errors and r squared results. I did not include the performances of the various neural networks I've used, but they generally fall somewhere right around the gradient boosting methods. I chose the extra trees regressor for fine tuning in the end, which is a subset of ensemble decision trees algorithms. The feature importance plot here for the extra trees regressor shows a logical assignment of feature importances, where things such as location and square footage of the properties are given priorities. The plot on the right here, shows a distribution of the median error of the estimates in the testing phase of the modelling. You can see here that it follows a normal distribution, with most of the error percentages congregating around 0.

Here is the result. A median absolute error of fifty one thousand, six hundred and eighty six dollars. Considering that the average property value in Vancouver is 1.2 million, I am pretty happy with the result given the time and resource constraint. It is also an added bonus that the model seems to outperform the industry standard for this, Zillow zestimate, by more than 2 percent. It is also worth noting that Zillow in fact does not offer the zestimate service for any Vancouver properties, perhaps due to the uniqueness of this market.

Now I would like to give a very quick demonstration for my estimate application in its proof-of-concept stage. Please click on the link that I've mentioned before the start of the presentation if you would like to give it a try afterwards.
There are mainly two parts for the app. The first is a demonstration of its ability to do batch prediction and display some basic data visualizations on the testing dataset. I can select the subset of the properties I want to predict on here.
Underneath this part is where you can input custom data for a property: there are such things as location, neighborhood, and square footage of the floor area and lot size. Once you are happy with your selection, click on this button here and the model will return you its prediction.

In order to take the application past its current proof-of-concept phase, the next things I will like to focus on are the following:

- Getting more data, that can properly account for the edge cases in Vancouver
- Explore more neural network approaches, with more hardware and time allowances
- Using a more robust server for my application
- Collaborating with web developers to make the app more visually appealing

Due to the time constraint of the presentation, there are many interesting details and findings that I had to leave out, such as how it is possible to write an advertisement that can potentially actually lower the estimated price of your property, based on your choice of syntax and sentence structures. Or the fact that the local indigenous history might affect land values in specific neighborhoods. If you have any questions, please don't hesitate to find me in the breakout session after the presentation. Thank you very much again for your time!