Columbia University - Data Analytics & Visualization Bootcamp

Group Kick-off 11/10/22

Proposed Agenda

- Introductions
- Project Objective
- Timeline and Project Segments
- Rubric
- Administration Items
- Next Steps

Team

- Adam Karim
- Deanna Centi
- Hande Demir-Neilan
- Madeline Propis
- Stefani Centi
- Vince Trombetta

Project Objective

This Photo by Unknown

The goal of this project is to develop a set of high-quality residential real estate sales leads that predict which Home Owners are considering selling their home, their motivations for doing so, and the timing of their actions. The geographic market for developing these leads is New Castle County, Delaware (pop. 550,000). Time period evaluated is TBD.

Timeline and Project Segments

Due Date	Description / Segments	% of Final Grade
12/07	Module 19 is completed (Neural Networks and Deep Learning)	
	Module 20 begins (Final Group Project)	
12/14	Segment 1 - “Introduction to the Group Project”	19%
12/21	Segment 2 – “Build the Pieces”	19%
12/26	<< Break >>	
01/04	Segment 3 – “Plug It In”	19%
01/11	Segment 4 – “Put It all Together”	40%
	Individual Self-Assessment	3%
TBD	Module 21 – Course Wrap Up	

Rubric
12/14/22 – “Introduction to the Group Project”

Presentation	GitHub	Machine Learning Model	Database	Dashboard
Team Members have drafted their project, including the following: ✓ Selected topic ✓ Reason why they selected their topic ✓ Description of their source of data ✓ Questions they hope to answer with the data *Note: The content does not yet need to be in the form of a presentation; text in the README.md works as well.*	**Main Branch** - Includes a README.md that must include: ✓ Description of the communication protocols **Individual Branches -** ✓ At least one branch for each team member ✓ Each team member has at least four commits from the duration of the first segment *Note: The descriptions and explanations required in all other project deliverables should also be in your README.md as part of your outline, unless otherwise noted.*	**Team Members** present a provisional machine learning model that stands in for the final machine learning model and accomplishes the following: ✓ Takes in data in from the provisional database ✓ Outputs label(s) for input data	**Team Members** present a provisional database that stands in for the final database and accomplishes the following: ✓ Sample data that mimics the expected final database structure or schema ✓ Draft machine learning module is connected to the provisional database	**n/a**

- Rubric based on demonstrated “Proficiency”

Rubric
12/21/22 – "Build the Pieces"

Presentation	GitHub	Machine Learning Model	Database	Dashboard
The presentation outlines the project, including the following: ✓ Selected topic ✓ Reason why they selected their topic ✓ Description of their source of data ✓ Questions they hope to answer with the data ✓ Description of the data exploration phase of the project ✓ Description of the analysis phase of the project *Note: Presentations are drafted in Google Slides.*	**Main Branch** - All code in the main branch is production- ready. Main branch should include: ✓ All code necessary to perform exploratory analysis ✓ Some code necessary to complete the machine learning portion of the project **README.md** must include: ✓ Description of the communication protocols ✓ Outline of the project (this may include images, but should be easy to follow and digest) *(continued next page)*	**Team** members submit the code for their machine learning model, as well as the following: ✓ Description of preliminary data preprocessing ✓ Description of preliminary feature engineering and preliminary feature selection, including their decision-making process ✓ Description of how data was split into training and testing sets ✓ Explanation of model choice, including limitations and benefits	**Team Members** present a fully integrated database: ✓ Database stores static data for use during the project ✓ Database interfaces with the project in some format (e.g., scraping updates the database, or database connects to the model) ✓ Includes at least two tables (or collections, if using MongoDB) ✓ Includes at least one join using the database language (not including any joins in Pandas) ✓ Includes at least one connection string (using SQLAlchemy or PyMongo) *(continued next page)*	**A blueprint** for the dashboard is created and includes all of the following: ✓ Storyboard on Google Slide(s) ✓ Description of the tool(s) that will be used to create final dashboard ✓ Description of interactive element(s)

Rubric
12/21/22 – “Build the Pieces” continued...

Presentation	GitHub	Machine Learning Model	Database	Dashboard
	Note: The descriptions and explanations required in all other project deliverables should also be in your README.md as part of your outline, unless otherwise noted. **Individual Branches -** ✓ At least one branch for each team member ✓ Each team member has at least four commits for the duration of the second segment (eight total commits per person)		*Note: If you use a SQL database, you must provide your ERD with relationships.*	

Rubric
01/04/23 – "Plug It In"

Presentation	GitHub	Machine Learning Model	Database	Dashboard
The presentation outlines the project, including the following: ✓ Selected topic ✓ Reason why they selected their topic ✓ Description of their source of data ✓ Questions they hope to answer with the data ✓ Description of the data exploration phase of the project ✓ Description of the analysis phase of the project ✓ Technologies, languages, tools, and algorithms used throughout the project *Note: Presentations are drafted in Google Slides.*	**Main Branch** - All code in the main branch is production- ready. Main branch should include: ✓ All code necessary to perform exploratory analysis ✓ Most code necessary to complete the machine learning portion of the project **README.md** must include: ✓ Description of the communication protocols has been removed. ✓ Cohesive, structured outline of the project (this may include images, but should be easy to follow and digest) *(continued next page)*	**Team** members submit the code for their machine learning model, as well as the following: ✓ Description of data preprocessing (replaces preliminary) ✓ Description of feature engineering and the feature selection, including their decision- making process (replaces preliminary) ✓ Description of how data was split into training and testing sets ✓ Explanation of model choice, including limitations and benefits ✓ Explanation of changes in model choice (if changes occurred between the Segment 2 and Segment 3 deliverables) *(continued next page)*	**Team members** present a final project with a fully integrated database. ✓ Database stores static data for use during the project ✓ Database interfaces with the project in some format (e.g., scraping updates the database, or database connects to the model) ✓ Includes at least two tables (or collections, if using MongoDB) ✓ Includes at least one join using the database language (not including any joins in Pandas) ✓ Includes at least one connection string (using SQLAlchemy or PyMongo) *(continued next page)*	**The dashboard** presents a data story that is logical and easy to follow for someone unfamiliar with the topic. It includes all of the following: ✓ Images from the initial analysis ✓ Data (images or report) from the machine learning task ✓ At least one interactive element

Rubric
01/04/23 – “Plug It In” continued...

Presentation	GitHub	Machine Learning Model	Database	Dashboard
	README.md must include: ✓ Link to Google Slides draft presentation *Note: The descriptions and explanations required in all other project deliverables should also be in your README.md as part of your outline, unless otherwise noted.* **Individual Branches:** ✓ At least one branch for each team member ✓ Each team member has at least four commits for the duration of the third segment (12 total commits per person)	✓ Description of how they have trained the model thus far, and any additional training that will take place ✓ Description of current accuracy score Additionally, the model obviously addresses the question or problem the team is solving.	Note: If you use a SQL database, you must provide your ERD with relationships.	

Rubric
01/11/23 – "Put It All Together"

Presentation	GitHub	Machine Learning Model	Database	Dashboard
The presentation outlines the project, including the following: ✓ Selected topic ✓ Reason why they selected their topic ✓ Description of their source of data ✓ Questions they hope to answer with the data ✓ Description of the data exploration phase of the project ✓ Description of the analysis phase of the project ✓ Technologies, languages, tools, and algorithms used throughout the project ✓ Result of analysis ✓ Recommendation for future analysis *(continued next page)*	**Main Branch** - All code in the main branch is production- ready. All code is clean, commented, easy to read, and adheres to a coding standard (e.g., PEP8). **Main Branch** should include: ✓ All code necessary to perform exploratory analysis ✓ All code necessary to complete machine learning portion of project ✓ Any images that have been created (at least three) ✓ Requirements.txt file *(continued next page)*	**Team** members submit the code for their machine learning model, as well as the following: ✓ Description of data preprocessing ✓ Description of feature engineering and the feature selection, including their decision- making process ✓ Description of how data was split into training and testing sets ✓ Explanation of model choice, including limitations and benefits ✓ Explanation of changes in model choice (if changes occurred between the Segment 2 and Segment 3 deliverables) *(continued next page)*	**Team members** present a final project with a fully integrated database. ✓ Database stores static data for use during the project ✓ Database interfaces with the project in some format (e.g., scraping updates the database, or database connects to the model) ✓ Includes at least two tables (or collections, if using MongoDB) ✓ Includes at least one join using the database language (not including any joins in Pandas) ✓ Includes at least one connection string (using SQLAlchemy or PyMongo) *(continued next page)*	**The dashboard** presents a data story that is logical and easy to follow for someone unfamiliar with the topic. It includes all of the following: ✓ Images from the initial analysis ✓ Data (images or report) from the machine learning task ✓ At least one interactive element Either the dashboard is published or the submission includes a screen capture video of it in action.

Rubric
01/11/23 – "Put It All Together" continued…

Presentation	GitHub	Machine Learning Model	Database	Dashboard
✓ Anything the team would have done differently **Google Slides:** ✓ Slides are primarily images or graphics (rather than primarily text) ✓ Images are clear, in high-definition, and directly illustrative of subject matter **Live Presentation:** ✓ All team members present in equal proportions ✓ The team demonstrates interactivity of dashboard in real time ✓ The presentation falls within any time limits provided by instructor ✓ Submission includes speaker notes, flashcards, or a video of the presentation rehearsal	**README.md** must include: ✓ Cohesive, structured outline of the project (this may include images, but should be easy to follow and digest) ✓ Link to dashboard (or link to video of dashboard demo) ✓ Link to Google Slides presentation *Note: The descriptions and explanations required in all other project deliverables should also be in your README.md as part of your outline, unless otherwise noted.* *(continued next page)*	✓ Description of how model was trained (or retrained, if they are using an existing model) ✓ Description and explanation of model's confusion matrix, including final accuracy score Additionally, the model obviously addresses the question or problem the team is solving. *Note: If statistical analysis is not included as part of the current analysis, include a description of how it would be included in the next phases of the project.*	Note: If you use a SQL database, you must provide your ERD with relationships.	

Rubric
01/11/23 – “Put It All Together” continued...

Presentation	GitHub	Machine Learning Model	Database	Dashboard
	Individual Branches: ✓ At least one branch for each team member ✓ Each team member has at least four commits for the duration of the final segment (16 total commits per person)			

Administrative Items

- Select a Group Name
- Determine when and how often to meet
- Discuss who wants to do what?
- Data Resources – mini-brainstorm
- Understand that subject matter for Machine Learning has not been taught yet

Next Steps

Appendix

Examples of Hypotheses

Only 1-2 can be undertaken due to project time constraints

Hypothesis	Data Sources / Considerations
Owners with children entering school will consider moving to get to a better school district; they must have equity in their current home in order to do so.	Demographics, birth records, mortgage information, school district ratings.
Owners at retirement age will consider selling to move to a retirement community or destination.	Demographics, regional migration patterns, current home size, availability and price of homes at destination, Zillow APIs
Changes in interest rates will motivate or inhibit owners to sell – both from a Buyer demand and purchase opportunity perspective.	Interest rate history and forecasts, correlated to past home sales; attempt to isolate effect
Single owners getting married will sell, respectively, then purchase a new home in joint tenancy	Demographics, marriage records, mortgage information
Changes in crime patterns will motivate owners to sell	Crime maps, historical trends, neighbor insights

Notes from Discussion w/ TAs

- Scope needs to be sufficient for six team members, but also executable within time frame
- Project probably requires Classification Logistic regression or Random Forest Classification
- Can use decision trees
- Look for balanced, unbiased distributions
- Watch out for over-fitting of model

Understanding the problems

01 A consistent source of quality leads is a challenge – Some of the best lead sources for agents are not always the most consistent as far as lead volume goes. For example, word-of-mouth, repeat business and referral networks usually produce the best leads, however, often there are simply not enough of them. And, while agents can undertake lots of activities to promote themselves within their network, it is a finite set of opportunities, especially for newer agents who are building a book-of-business. What agents need are lead sources that supplement their personal sphere, especially if the goal is to grow.

Understanding the problems

02

Sellers are more elusive than Buyers– Over 80% of home buyers use the internet to research, shop, compare and narrow down their choices, including those buyers who are actively working with an agent. This gives sites such as Zillow, Redfin, Trulia and Realtor.com an opportunity to capture information on who is using their sites and sell these buyer leads to agents. On the other hand, Sellers may use the internet to compare their home to others, but they do not “look for buyers” on the internet, unless it is for-sale-by-owner, a fraction of all listings. Other ways are needed to predict who might be selling a home, why they are selling, and when.

Understanding the problems

03

Multiple sources of data need to be connected – There are many potential sources of data that could be useful in predicting who might be selling their home – public records, mortgage data, tax records, census data, demographic data, the multi-listing service, neighbourhood ratings, school-district scores, crime maps, api's for web services like Zillow, purchased data such as core logic, etc. The challenge is putting it all together to develop an ecosystem that predicts just who is thinking about selling, why they are selling, and when they will do so. And, while there are players in this space, that is, those who develop and provide seller leads for a fee, the accuracy of those leads is a challenge and no one seems to have 'cracked the code'. One of the competitive advantages is having access to actual agents and a broker to add a layer of qualitative data, which we will have in the course of this project.