Supplemental Material: Slides for second interview study

Evaluating Regression Models

Demonstration with Subject Matter Experts

PI: redacted

PI: redacted

Interview Agenda

- The interview will not go past the hour we have
- 3 parts in the interview, where you will be assessing two new regression models & their performance
    - 10 minutes for Part 1 (Scenario 1, Dataset #1, Regression Model #1)
    - 30 minutes for Part 2 (Scenario 2, Dataset #2, Regression Model #2, Extra slides)
    - 10 minutes for Q&A (Wrap-up)
- This interview will be transcribed and possibly published in research
    - Feel free at any time to speak in abstractions
    - *redacted* and I will remove all sensitive information and anonymize all quotes
    - If you'd like to provide further comments or remove any comments after the interview, please let us know at any time
- Any questions before we continue?

Preparation

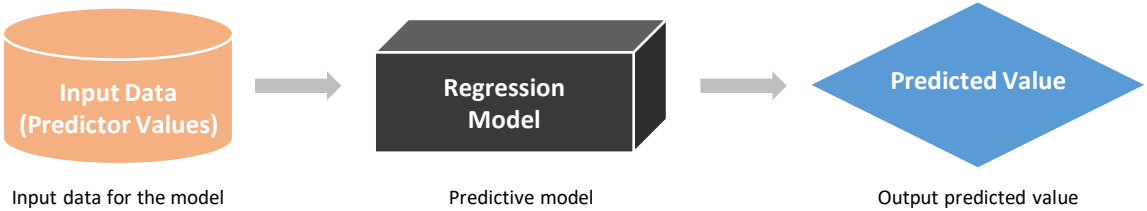
You are participating in a study as a subject matter expert who interacts with regression models. We will walk you through two different theoretical scenarios (Scenario 1, Scenario 2). Afterwards, we we will ask you follow-up questions for both scenarios.

For each scenario, we will introduce an associated dataset, prediction task, and regression model that is currently in-development. We will discuss the quality of each regression model with you, and at the end we will ask you to assess whether you think the new model should be used for the described dataset and prediction task.

It is important to note that there is **absolutely no right or wrong answer**. The purpose of this interview is to *better understand which components of a presentation best explain the performance of a regression model*.

Refresher on Regression Models

- Similar to our first interview, when we talk about a regression model, we refer to any predictive model that takes in data and outputs a predicted numerical value
  - Unlike classification models, which output a single ‘yes’ or ‘no’ decision, regression models output a predicted number (e.g., ‘7’ or ‘0.023’)
- A classic example of a regression model is one used at a weather station to predict the daily temperature based on a set of weather conditions as input
- Any questions about regression models before we continue?



Scenario **(1)**

Dataset: Automobiles & their MPG

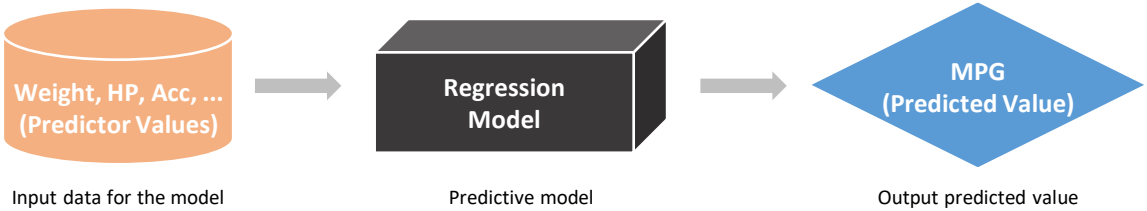
Automobiles & their MPG Regression Model

Dataset: We are working with a ‘automobiles’ dataset which contains information about a number of vehicles.

Prediction Task: The in-development regression model attempts to predict a vehicle’s *MPG* based on the vehicle’s attributes.

- MPG (miles per gallon) is the distance, measured in miles, that a vehicle can travel per gallon of fuel. MPG is also the primary measurement of a vehicle’s fuel efficiency: The higher a vehicle’s MPG, the more fuel efficient it is.

Your Task: You work for a car manufacturer and you would like to predict the MPG of a car you’re designing *before* you actually build it. After this presentation, you will decide whether you think this model can help you in assessing a vehicle’s MPG before it is built.



Description of the problem & data

Prediction Task: Use a regression model to predict a vehicle's MPG.

Model Input: A set of data attributes related to a vehicle.

- MPG
- Displacement
- Horsepower
- Weight
- Acceleration
- Model year

Model Output / Target: A numerical value (##.#) representing the vehicle's MPG.

Dataset example

MPG	Displacement	Horsepower	Weight	Acceleration	Model year
30.7	145	76	3,160	19.6	81
18	250	78	3,574	21	76
21	122	86	2,226	16.5	72
19	225	100	3,630	17.7	77
18	225	105	3,121	16.5	73
29	135	84	2,525	16	82

Each row represents a unique automobile. Each column is an attribute in the dataset.
The target (attribute the model is predicting) is MPG.
The predictors (inputs to the model) are Displacement, Horsepower, Weight, Acceleration, and Model year.

Model type & performance

We will now shift towards presenting the in-development model's performance. To further illuminate the quality of the new model, we will compare it to a baseline model. We can assume the baseline model is *already in use*, and we are assessing whether we should replace the baseline model with the new model. For reference, the in-development model is an "XGB" model, and the baseline / standard model is a "KNN" model.

Baseline Model: K-nearest neighbors (KNN)

New Model: Extreme gradient boosting decision tree (XGB)

KNN vs. XGB metric performance

Model	R2	MSE	MAE
KNN (baseline)	.84	8.36	2.22
XGB (in-development)	.88	6.51	1.95

Examples of results

KNN (baseline)	XGB (new)	MPG	Displacement	Horsepower	Weight	Acceleration	Model year
26.129	27.846	30.7	145	76	3,160	19.6	81
18.621	18.725	18	250	78	3,574	21	76
24.714	24.224	21	122	86	2,226	16.5	72
19.307	19.58	19	225	100	3,630	17.7	77
18.286	17.039	18	225	105	3,121	16.5	73
30.829	30.241	29	135	84	2,525	16	82

Data scientist’s conclusion

How it relates to business objectives / original prediction task:

The XGB model has better R2, MSE, and MAE than the KNN model. This means that it should more accurately predict MPG for a car design.

Scenario **(2)**

Dataset: Housing & their values

In scenario 2, we will walk through a different dataset and regression task. This time, we will show extra information about the models. Afterwards, we will ask you about the effectiveness of the extra information.

Slides with extra information compared to scenario 1 will have the following symbol: Extra

Extra

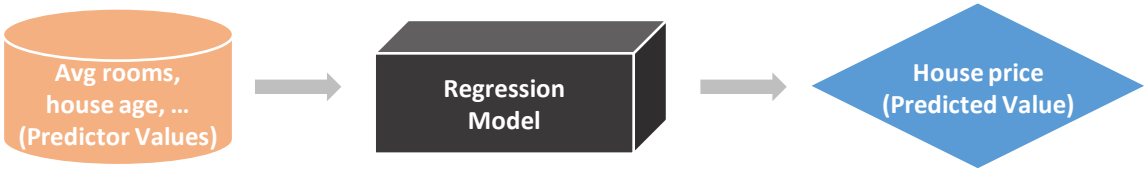
At any time, we encourage you to **stop us to ask any clarifying questions**. This includes questions about the data used for training the model, questions about any of the charts we will show you, or questions about any of the language or vocabulary we use. If you find that we are presenting concepts too slowly, please feel free to ask us to speed up.

California housing & their prices Regression Model

Dataset: We are working with a ‘housing’ dataset which contains information about a number of houses and their values in California, USA.

Prediction Task: The in-development regression model attempts to predict a California house’s *value* based on the house’s attributes.

Your Task: You work for a California real estate firm and want to predict the potential value of a home before deciding to invest in its renovations and rebuilds. After this presentation, you will decide whether you think this model could help you in assessing a house’s value before you further invest in it.



Description of the problem & data

Prediction Task: Use a regression model to predict a California house's value.

Model Input: A set of data attributes related to a house in California.

- House value
- Median income
- House age
- Population
- Avg. rooms
- Avg. bedrooms
- Avg. occupancy
- Latitude
- Longitude

Model Output / Target: A numerical value ($###,###) representing the house's price value.

Extra

Data descriptors

Data Attribute	Description
House value	House value (measured in US Dollars)
Median income	Median income for households within a block of houses (measured in tens of thousands of US Dollars)
House age	The age of the house
Population	Total number of people residing within a block of houses
Avg. rooms	Average number of rooms within a house
Avg. bedrooms	Average number of bedrooms within a house
Avg. occupancy	Average number of occupants within a house
Latitude	A measure of how far north a house is; a higher value is farther north
Longitude	A measure of how far west a house is; a higher value is farther west

Dataset example

Value	Med income	House age	Population	Avg rooms	Avg bedrooms	Avg occupancy	Latitude	Longitude
47,700	1.681	25	1,392	4.192	1.022	3.877	36.06	-119.01
45,800	2.531	30	1,565	5.039	1.193	2.68	35.14	-119.46
500,001	5.738	52	1,310	3.977	1.186	1.36	37.8	-122.44
218,600	3.725	17	1,705	6.164	1.028	3.444	34.28	-118.72
158,700	4.715	34	1,063	5.493	0.975	2.484	36.62	-121.93
198,200	5.084	12	2,400	5.251	1.095	2.847	34.08	-117.61

Each row represents a unique house. Each column is an attribute in the dataset.
The target (attribute the model is predicting) is a house's Value.
The predictors (inputs to the model) are Med Income, House age, Population, Avg rooms, Avg bedrooms, Avg occupancy, Latitude, and Longitude.

Extra

Model type & performance

We will now shift towards presenting the in-development model's performance. To further illuminate the quality of the new model, we will compare it to a baseline model. We can assume the baseline model is *already in use*, and we are assessing whether we should replace the baseline model with the new model. For reference, the in-development model is an "XGB" model, and the baseline / standard model is a "KNN" model.

Baseline Model: K-nearest neighbors (KNN) looks for the *K* most similar examples and takes the average value for them as its prediction.

New Model: Extreme gradient boosting decision tree (XGB) builds many tree-based models on subsets of rows and columns of the data and averages out their predictions.

Extra

KNN vs. XGB performance

Model	r2	MSE	MAE
KNN (baseline)	.7	3,956,803,800	42,271.09
XGB (in-development)	.83	2,270,764,000	31,662.65

- **XGB has a higher r2 than KNN**. r2 measures the proportion of the total variance explained by the model to the total variance. A higher r2 generally means the model better fits the data.
- **XGB has a lower MSE than KNN**. MSE measures the average of the square of the model's errors. A lower MSE generally means a better forecast as the errors are smaller.
- **XGB has a lower MAE than KNN**. MAE measures the average magnitude of the errors in a set of predictions. A lower MAE indicates better model accuracy.

Extra

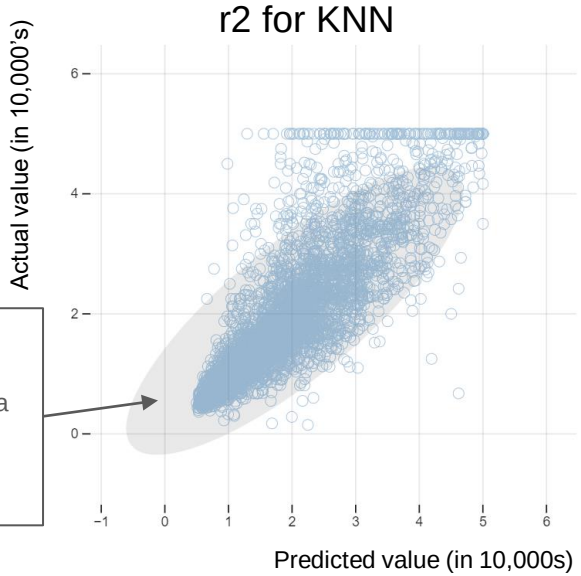
Visualizations for model performance

We will now present a series of visualizations that convey the performance of both the KNN and XGB models.

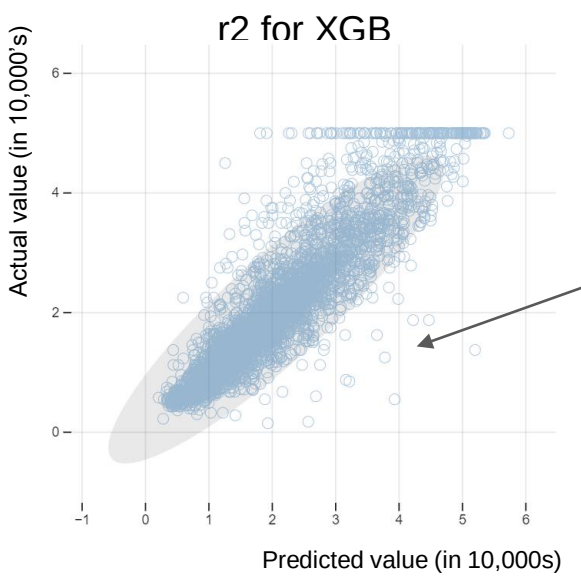
Please stop us if there is anything you are confused about.

Correlation scatterplot

Extra



KNN has a larger ellipse than XGB, signifying that it has a lower r2 and thus worse predictions on average.



Although XGB has a higher density of points contained within the 95% confidence ellipse, it is still susceptible to outliers. These points are where the model is predicting houses to have a higher value than they actually do.

- Higher r2 leads to skinnier ellipse - closer predictions on average
- For most houses, XGB predicts higher values than kNN

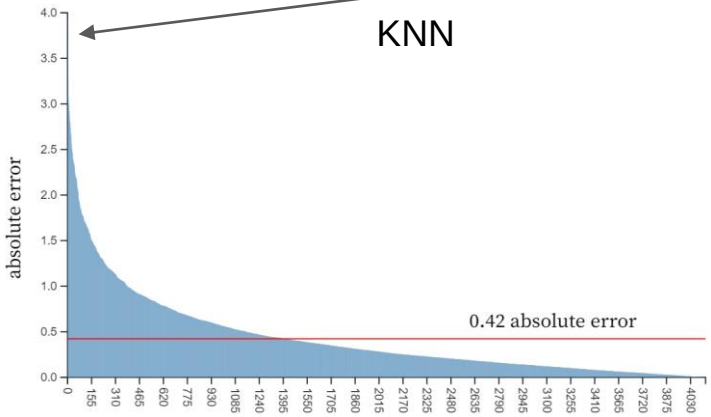
Bar chart with global absolute error line

Extra

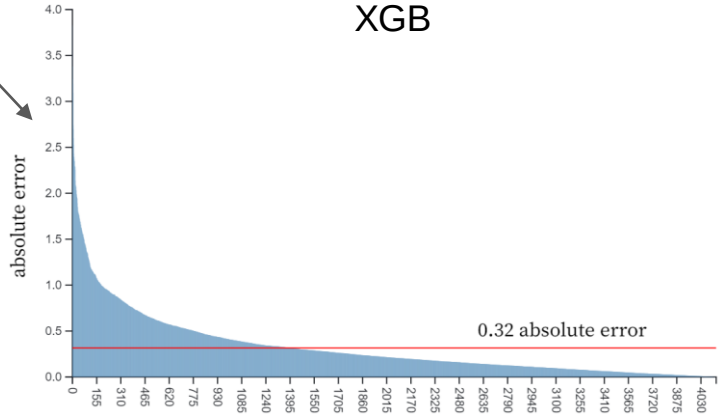
KNN has an average absolute error of 0.42, while XGB has a lower average absolute error of 0.32.

The highest absolute error on a single instance of KNN is 3.94, and the highest for XGB is 3.82.

KNN

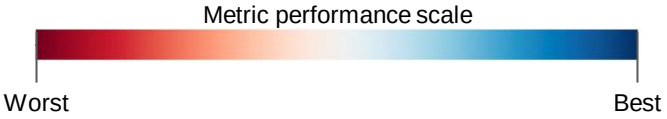


XGB



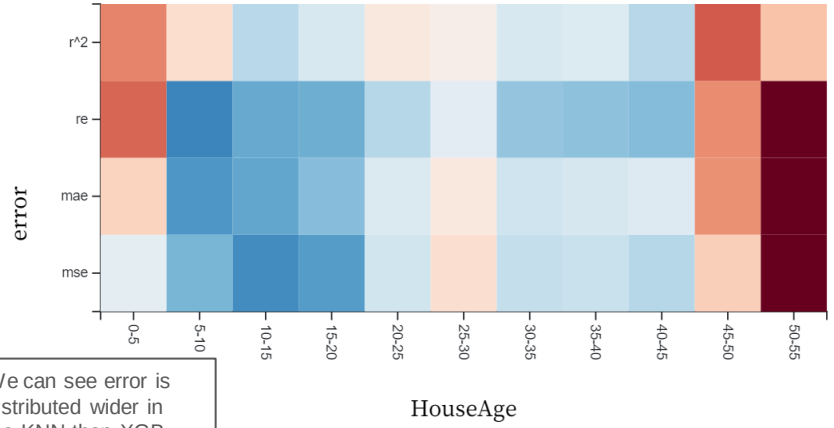
Heatmap by error category: House Age

Extra

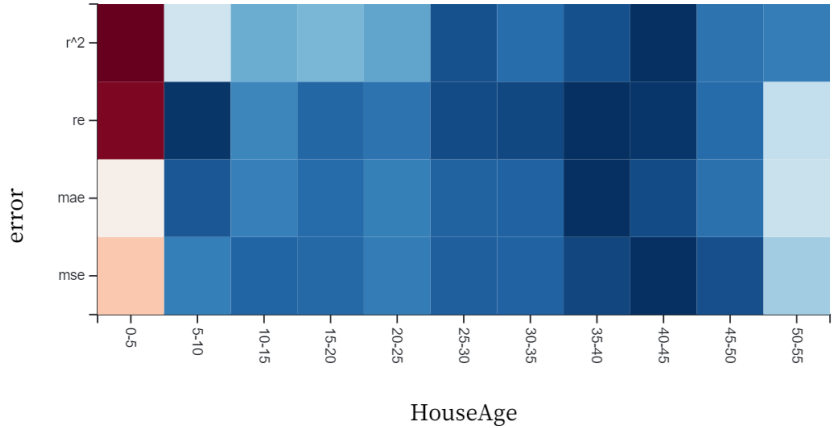


In general, XGB has lower error across all metrics than KGB for the attribute “House age”

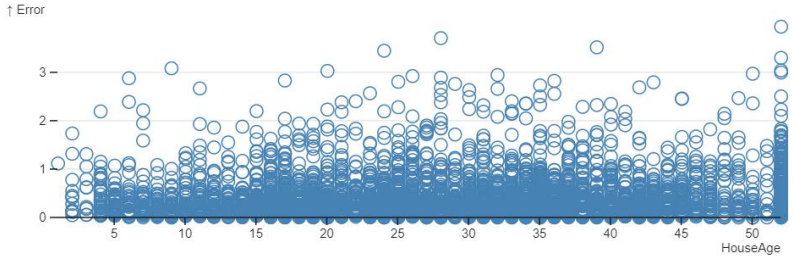
KNN

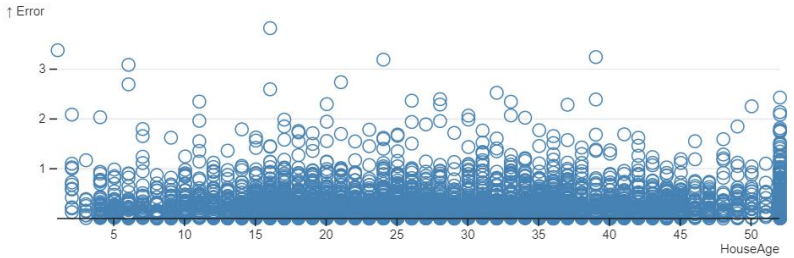


XGB



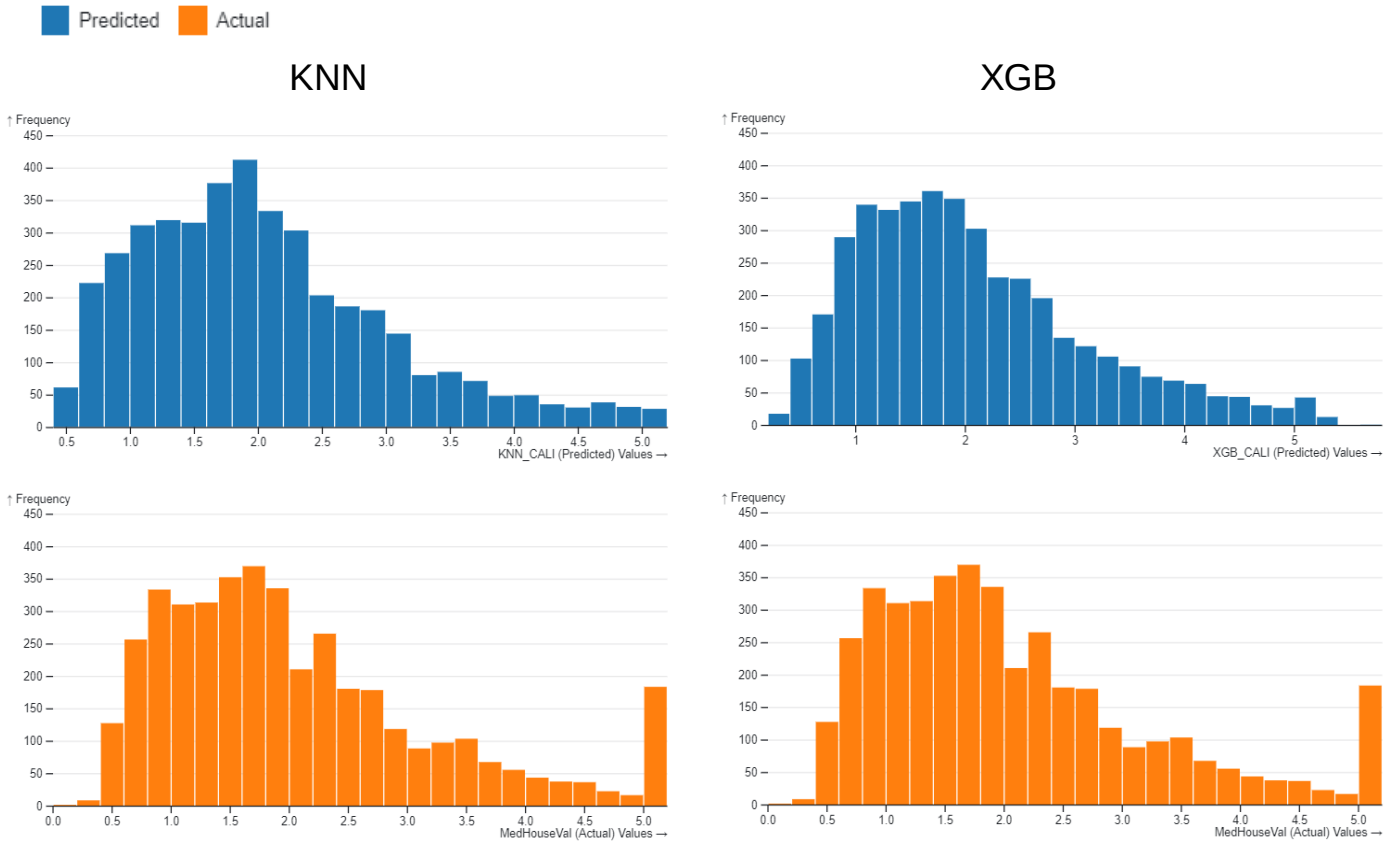
We can see error is distributed wider in the KNN than XGB model for House Age





Histograms for predicted vs. actual values

Extra

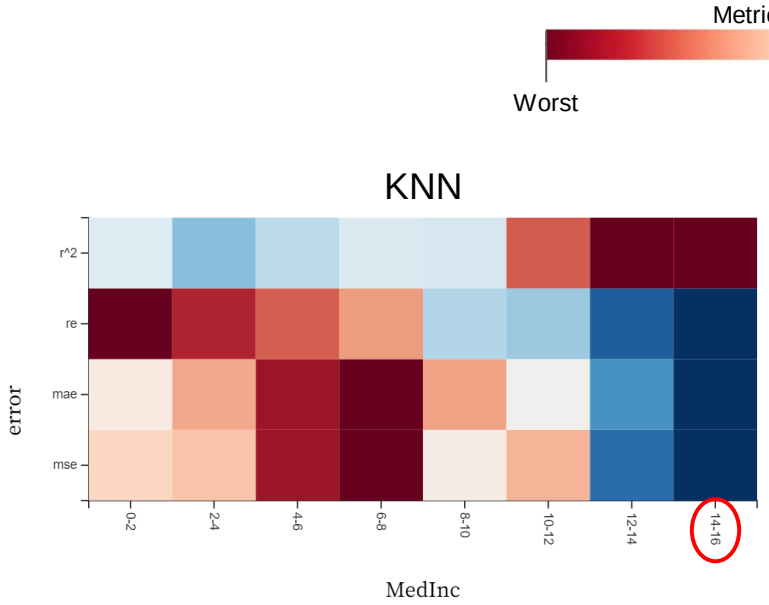


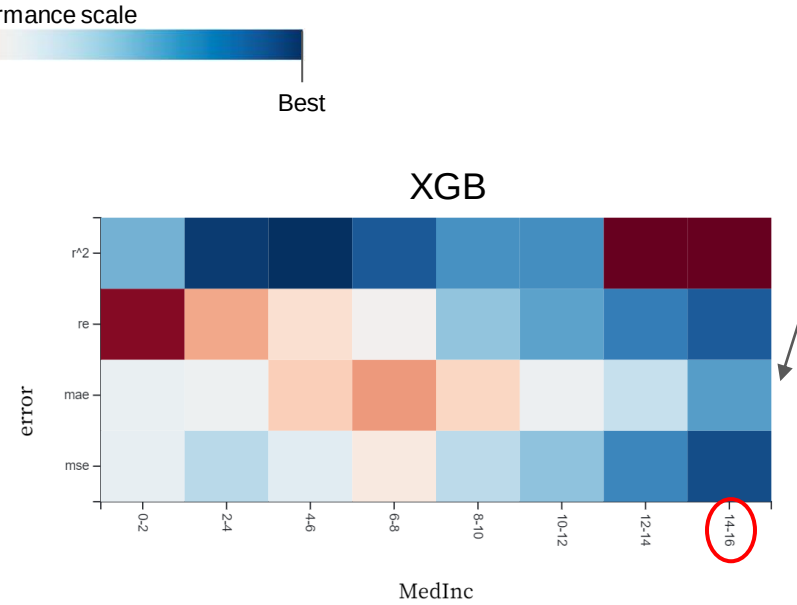
Difference in the shape of the models' predicted & actual value distribution

Large block on the right illustrates the models' difference in R2

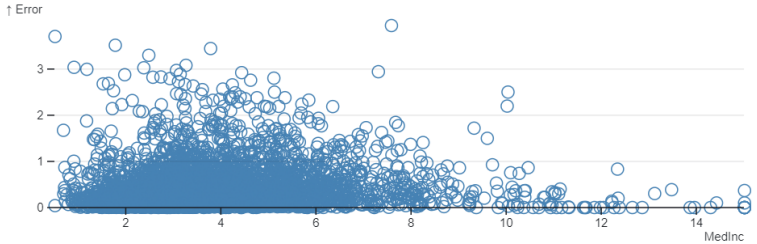
Examples of the model performing poorly

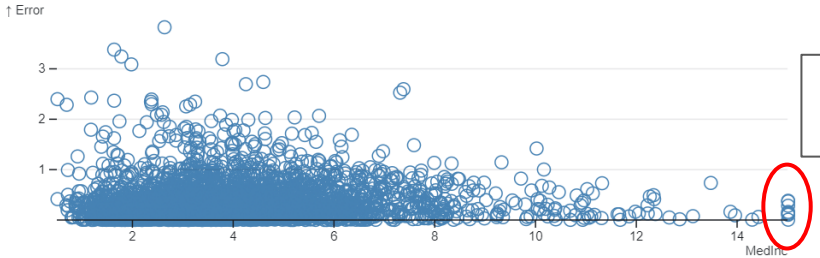
Extra





XGB does worse than KNN in predicting the highest Median Income subgroup. XGB may have a harder time predicting outliers where median income is the highest in the dataset.

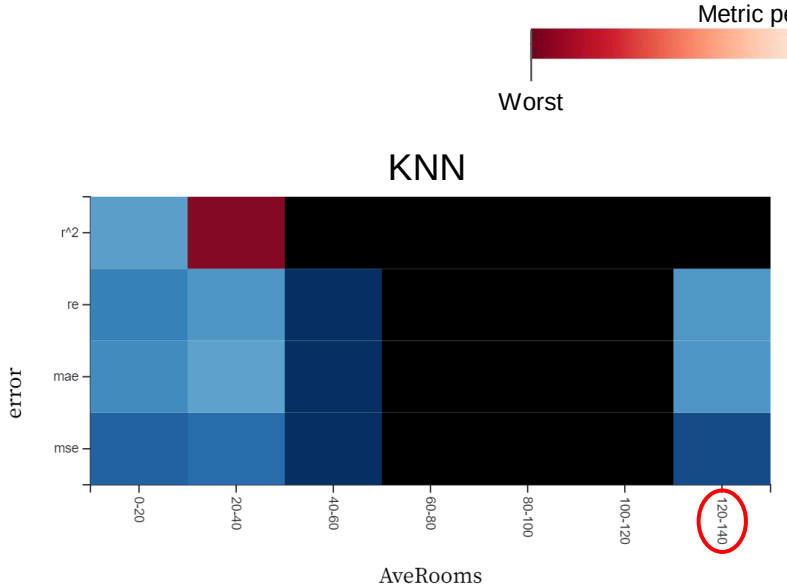


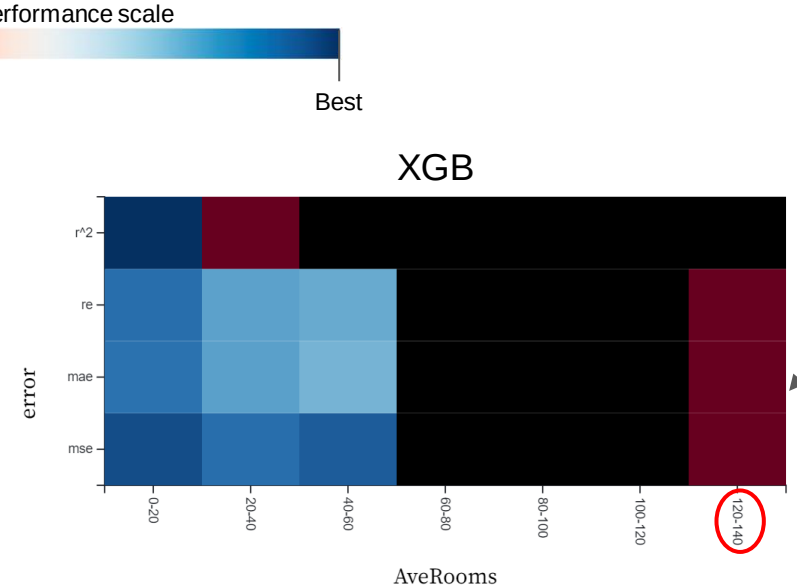


More dense outliers in XGB vs KNN

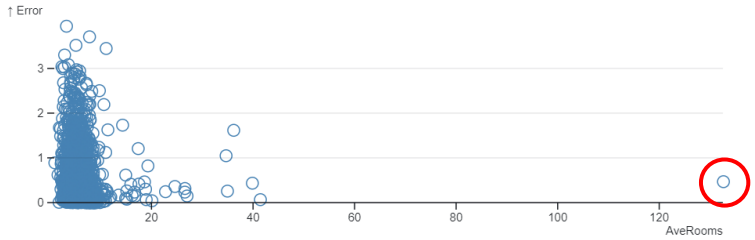
Examples of the model performing poorly

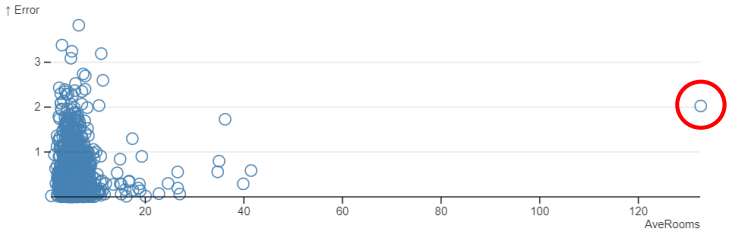
Extra





Difference in performance for houses where the # of avg rooms is between 120-140. XGB may be having a hard time predicting outliers where the average rooms is 120+





Some outliers have higher error in XGB than KNN

Extra

Examples of results

The model with the closer prediction to the target is highlighted

KNN (baseline)	XGB (new)	Value	Med income	House age	Population	Avg rooms	Avg bedrooms	Avg occupancy	Latitude	Longitude
55,800	58,315.77	47,700	1.681	25	1,392	4.192	1.022	3.877	36.06	-119.01
78,210	86,520.51	45,800	2.531	30	1,565	5.039	1.193	2.68	35.14	-119.46
432,430.3	**490,127.62**	500,001	5.738	52	1,310	3.977	1.186	1.36	37.8	-122.44
285,390	**261,883.53**	218,600	3.725	17	1,705	6.164	1.028	3.444	34.28	-118.72
249,750	**241,707.67**	158,700	4.715	34	1,063	5.493	0.975	2.484	36.62	-121.93
178,770	164,115.36	198,200	5.084	12	2,400	5.251	1.095	2.847	34.08	-117.61
211,720	239,165.62	198,200	3.691	38	4.963	1.048	1,011	3.758	33.92	-118.08
167,400	**147,873.56**	157,500	4.804	4	3.925	1.036	1,050	1.798	37.39	-122.08
224,860	**297,412.44**	340,000	8.113	45	6.879	1.012	943	2.782	34.18	-118.23

We can see that there are instances where XGB overestimates the value of a home more than KNN, although the overall predictions are more accurate. We believe the inaccuracies of XGB could be due to its inability to correctly predict large outliers in Median Income, Average rooms, and Average bedrooms.

Data Scientist’s conclusion

How it relates to business objectives / original prediction task:

The XGB model has better R2, MSE, and MAE than the KNN model. This means that it should more accurately predict the value of a house.

Wrap-up: Q&A

Follow up

We will now show you specific differences in our presentation from scenario 1 and scenario 2. We ask that you answer our questions **as honestly as possible.**

You will rate the effectiveness or helpfulness of each difference on a 7-point scale. but please feel free to add any more context or explanation in your response.

All differences from Scenario 2 vs. Scenario 1

- Offer to be stopped for clarification during presentation
- Data description table for all input to the model
- Descriptive explanation of the KNN and XGB models
- Definition of the performance metrics r2, MSE, MAE
- Visualizations showing comparisons of model performance
- Specific examples of XGB performing worse than KNN

As we walk you through these differences to rate their effectiveness one by one, please let us know if any one in particular was helpful or unhelpful to you.

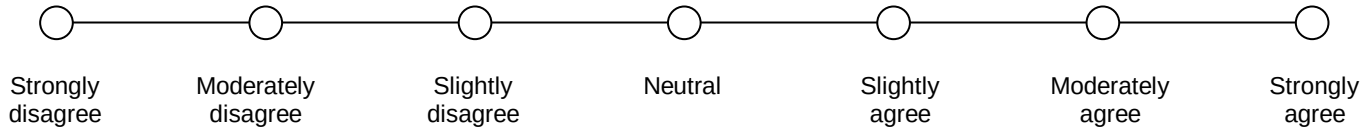
Difference 1: Offer to be stopped and/or sped up

At the start of Scenario 2, we asked you to stop us at any point to ask us clarifying questions about our presentation, charts, and vocabulary. In contrast, in Scenario 1, we did not mention this.

> At any time, we encourage you to **stop us to ask any clarifying questions**. This includes questions about the data used for training the model, questions about any of the charts we will show you, or questions about any of the language or vocabulary we use. If you find that we are presenting concepts too slowly, please feel free to ask us to speed up.

Please rate the following statement:

“I found this difference to be helpful in interpreting the model’s performance.”



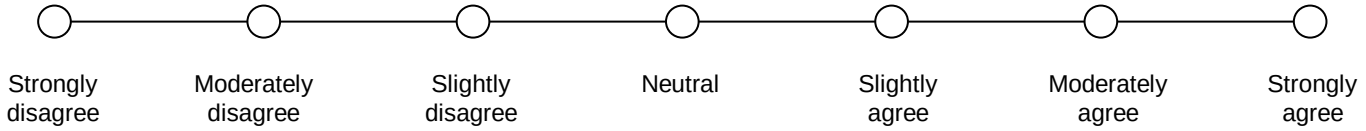
Difference 2: description of the input data

We provided data descriptions and units for the model in Scenario 2. In contrast, in Scenario 1, we only showed you the names of the attributes.

Data Attribute	Description
House value	House value (measured in US Dollars)
Median income	Median income for households within a block of houses (measured in tens of thousands of US Dollars)
House age	The age of the house
Population	Total number of people residing within a block of houses
Avg. rooms	Average number of rooms within a house
Avg. bedrooms	Average number of bedrooms within a house
Avg. occupancy	Average number of occupants within a house
Latitude	A measure of how far north a house is; a higher value is farther north
Longitude	A measure of how far west a house is; a higher value is farther west

Please rate the following statement:

“I found this difference to be helpful in interpreting the model’s performance.”



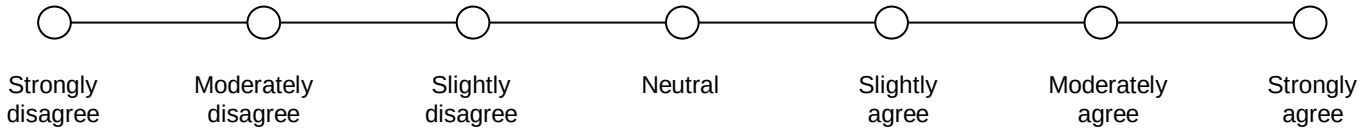
Difference 3: descriptive explanation of the models

We provided a short description for the regression models in Scenario 2. In contrast, in Scenario 1, we only showed you the names of the models.

> Baseline Model: K-nearest neighbors (KNN) looks for the *K* most similar examples and takes the average value for them as its prediction.
>
> New Model: Extreme gradient boosting decision tree (XGB) builds many tree-based models on subsets of rows and columns of the data and averages out their predictions.

Please rate the following statement:

“I found this difference to be helpful in interpreting the model’s performance.”



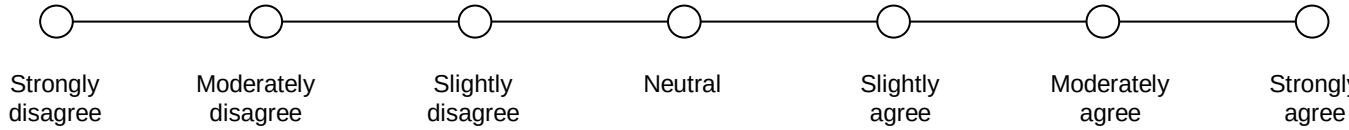
Difference 4: definition of performance metrics

We provided definitions of the metrics r2, MAE, and MSE in Scenario 2. In contrast, in Scenario 1, we only showed you the values for each metric.

- **XGB has a higher r2 than KNN**. r2 measures the proportion of the total variance explained by the model to the total variance. A higher r2 generally means the model better fits the data.
- **XGB has a lower MSE than KNN**. MSE measures the average of the square of the model's errors. A lower MSE generally means a better forecast as the errors are smaller.
- **XGB has a lower MAE than KNN**. MAE measures the average magnitude of the errors in a set of predictions. A lower MAE indicates better model accuracy.

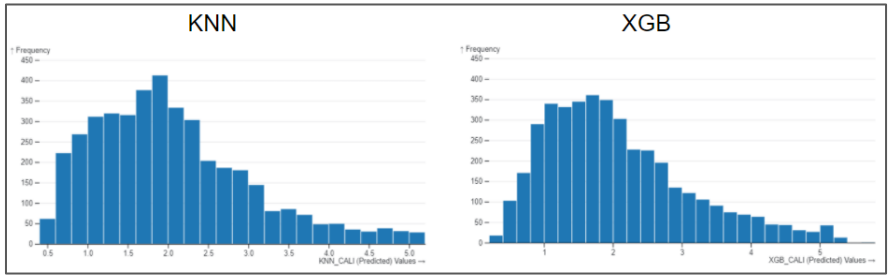
Please rate the following statement:

"I found this difference to be helpful in interpreting the model's performance."



Difference 5: visualizations with context

We provided a series of visualizations to convey the performance of our regression models in Scenario 2. In contrast, in Scenario 1, we only showed you the table of metrics.



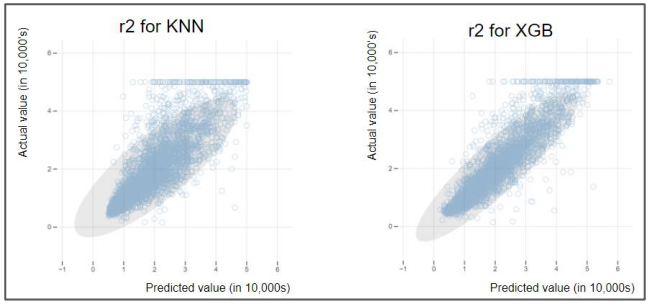
Please rate the following statement:

“I found this difference to be helpful in interpreting the model’s performance.”

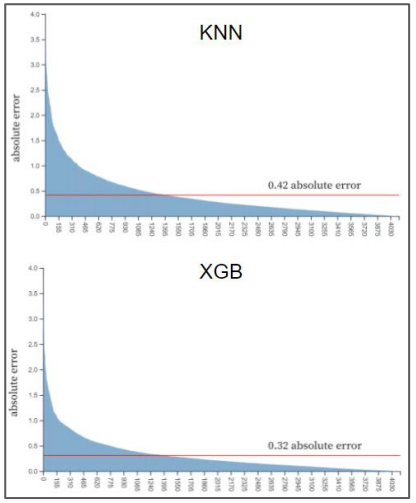
Strongly disagree | Moderately disagree | Slightly disagree | Neutral | Slightly agree | Moderately agree | Strongly agree

Difference 5: visualizations with context

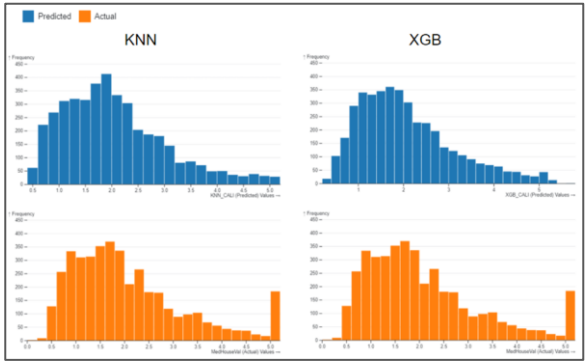
Of the visualizations we showed you, please tell us which were the most helpful and least helpful in interpreting a model's performance. If there was a visualization that was confusing to you, please indicate that to us as well.



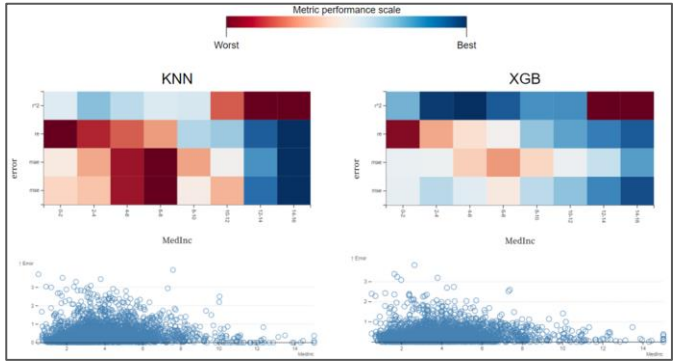
Correlation scatterplot



Global error barchart



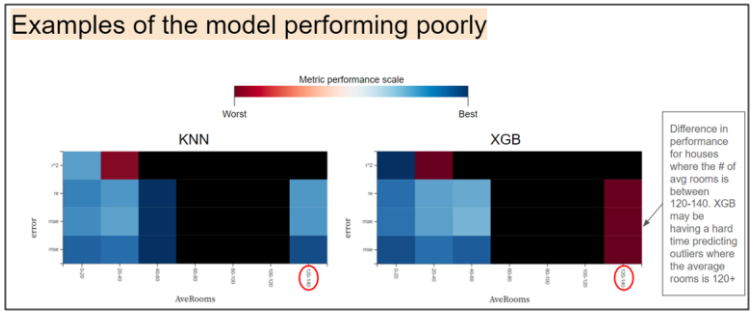
Histogram of error distribution



Heat map + Individual feature error plot

Difference 6: specific examples of the model performing poorly

We provided a few examples of the XGB model performing better and worse than the KNN model with our stipulations as to why in Scenario 2. In contrast, in Scenario 1, we did not provide any specific examples of the model.



The model with the closer prediction to the target is highlighted

We can see that there are instances where XGB overestimates the value of a home more than KNN, although the overall predictions are more accurate. We believe the inaccuracies of XGB could be due to its inability to correctly predict large outliers in Median Income, Average rooms, and Average bedrooms.

KNN (baseline)	XGB (new)	Value
55,800	58,315.77	47,700
78,210	86,520.51	45,800
432,430.3	490,127.62	500,001
285,390	261,883.53	218,600
249,750	241,707.67	158,700
178,770	164,115.36	198,200
211,720	239,165.62	198,200
167,400	147,873.56	157,500
224,860	297,412.44	340,000

Please rate the following statement:

“I found this difference to be helpful in interpreting the model’s performance.”

Strongly disagree | Moderately disagree | Slightly disagree | Neutral | Slightly agree | Moderately agree | Strongly agree

END OF THE INTERVIEW

Thank you again for participating and giving us your time today!

If at any time you would like to add further comments or remove comments from our interview, do not hesitate to reach out to *redacted* or me via email, teams chat, or call.

- *redacted* email
- *redacted* email

redacted and I will remove any sensitive information that might have come up during the interview. All information and quotes will be anonymized.

Do you have any other questions about our study or our interview?