DSIF5 Project 2 - Ames Housing

Team 2
Loh Wei Hao
Marc Tan
Ethan Lee
16 July 2022

Problem Statement

We are WEM general insurance based in Ames, Iowa specializing in **home insurance.**

Customers come to us looking to get their homes insured. A part of getting their homes insured requires the valuation of their property.

We have noticed through feedback forms that customers find our application forms:

- **Tedious and overly complicated**
- Usually **take more than an hour** for customers to fill up (total of 80 questions)
- Customers do not want to spend **more than 10 mins**

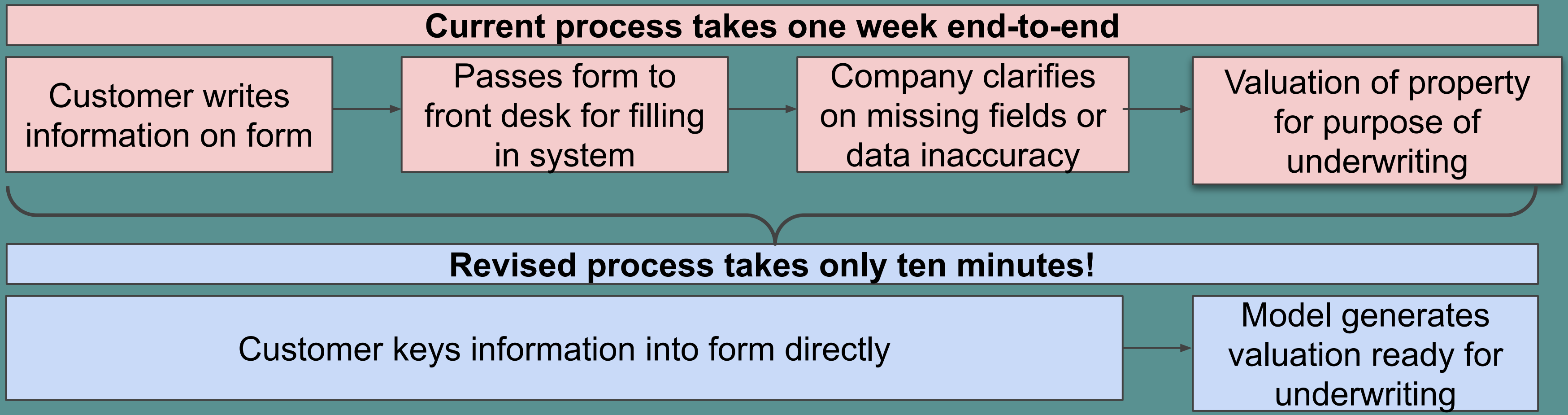
Due to the **high dropout rates**, management is concerned with the **loss of revenue and share of customers** to our competitors, who offer quicker and more accurate processing times.

Our immediate task

As a group of data scientists within the firm, we have been tasked to simplify the application process through **automating this process using machine learning models**.

These predictive valuation models will help to:

- **Effectively predict the valuation** of the property
- **Improve efficiencies and the overall customer end-to-end experience**
- Increasing take-up rate due to quicker turnaround times through quicker processing times



Current process takes one week end-to-end

Customer writes information on form → Passes form to front desk for filling in system → Company clarifies on missing fields or data inaccuracy → Valuation of property for purpose of underwriting

Revised process takes only ten minutes!

Customer keys information into form directly → Model generates valuation ready for underwriting

Our current application form

This is a sample of our form (Page 1 out of 5).

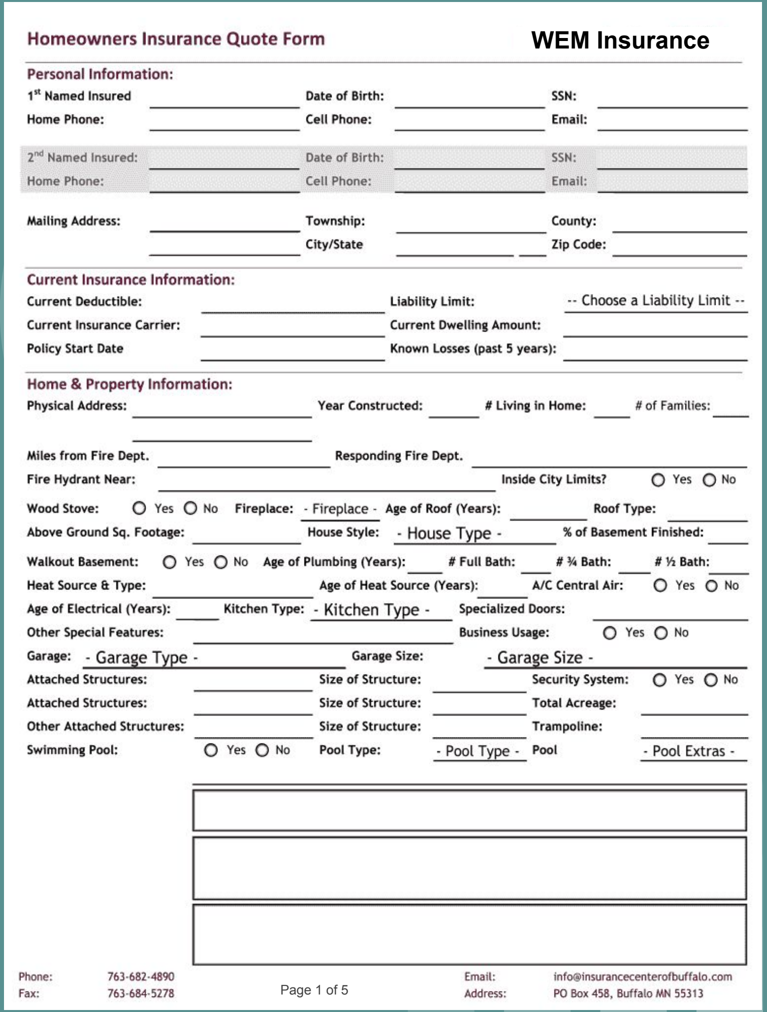
Customers tend to only complete one page and drop out.

Given that the form is **manual**, customers who make it to the end:

- Usually fill up the forms incorrectly (funny values)
- Leave plenty of blanks (null values)
- In turn, data inaccuracy and missing fields will require follow-ups (in turn increasing the whole on-boarding process)

When the front desk staff transfers this information to the database:

- Keys in the wrong values (human errors)
- Leave plenty of blanks (null values)



Homeowners Insurance Quote Form **WEM Insurance**

Personal Information:

1st Named Insured ______ Date of Birth: ______ SSN: ______
Home Phone: ______ Cell Phone: ______ Email: ______

2nd Named Insured: ______ Date of Birth: ______ SSN: ______
Home Phone: ______ Cell Phone: ______ Email: ______

Mailing Address: ______ Township: ______ County: ______
______ City/State ______ ______ Zip Code: ______

Current Insurance Information:

Current Deductible: ______ Liability Limit: -- Choose a Liability Limit --
Current Insurance Carrier: ______ Current Dwelling Amount: ______
Policy Start Date ______ Known Losses (past 5 years): ______

Home & Property Information:

Physical Address: ______ Year Constructed: ______ # Living in Home: ______ # of Families: ______
______
Miles from Fire Dept. ______ Responding Fire Dept. ______
Fire Hydrant Near: ______ Inside City Limits? ○ Yes ○ No
Wood Stove: ○ Yes ○ No Fireplace: - Fireplace - Age of Roof (Years): ______ Roof Type: ______
Above Ground Sq. Footage: ______ House Style: - House Type - % of Basement Finished: ______
Walkout Basement: ○ Yes ○ No Age of Plumbing (Years): ______ # Full Bath: ______ # ¾ Bath: ______ # ½ Bath: ______
Heat Source & Type: ______ Age of Heat Source (Years): ______ A/C Central Air: ○ Yes ○ No
Age of Electrical (Years): ______ Kitchen Type: - Kitchen Type - Specialized Doors: ______
Other Special Features: ______ Business Usage: ○ Yes ○ No
Garage: - Garage Type - Garage Size: - Garage Size -
Attached Structures: ______ Size of Structure: ______ Security System: ○ Yes ○ No
Attached Structures: ______ Size of Structure: ______ Total Acreage: ______
Other Attached Structures: ______ Size of Structure: ______ Trampoline: ______
Swimming Pool: ○ Yes ○ No Pool Type: - Pool Type - Pool - Pool Extras -

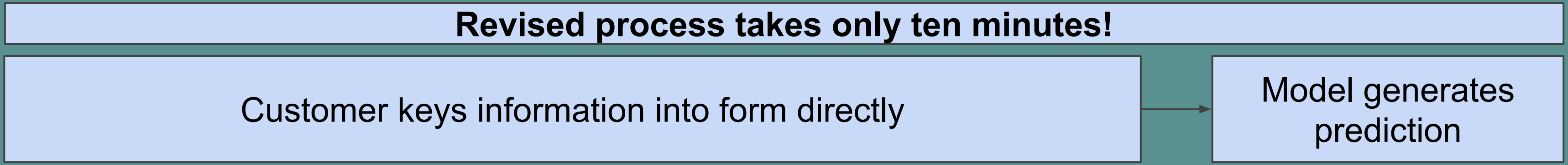
Phone: 763-682-4890
Fax: 763-684-5278

Page 1 of 5

Email: info@insurancecenterofbuffalo.com
Address: PO Box 458, Buffalo MN 55313

How we intend to improve the process

- **Eliminate manual forms (to save staff costs and time taken)**
- **Introducing online forms (to save paper and less answer errors)**
- **Greatly reduce the number of questions in form through data science!**
    - Customers must be able to complete under 10 mins
    - We aim for less than 20 questions
- **Combine similar or repeated questions**
    - No open ended questions
    - Customers would only be required to tick checkboxes, answer yes/no/others, or indicate based on a number scale (data validation checks front-end)



Revised process takes only ten minutes!

Customer keys information into form directly

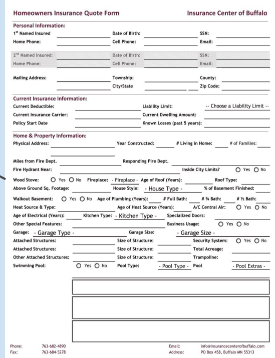
Model generates prediction

Database of historical sale prices

- Management has given us access to a historical database of 2051 homes in the Ames area.
- Dataset shows 80 sets of features as captured from the original application form.
- Dataset also shows the sale price of each home when it was sold. We will use this as the valuation of the homes.

X = 80 features
from the questions…

… from 2051 of
these forms!



Homeowners Insurance Quote Form **Insurance Center of Buffalo**

Personal Information:

1st Named Insured ____ Date of Birth: ____ SSN: ____
Home Phone: ____ Cell Phone: ____ Email: ____

2nd Named Insured: ____ Date of Birth: ____ SSN: ____
Home Phone: ____ Cell Phone: ____ Email: ____

Mailing Address: ____ Township: ____ County: ____
____ City/State ____ ____ Zip Code: ____

Current Insurance Information:

Current Deductible: ____ Liability Limit: -- Choose a Liability Limit --
Current Insurance Carrier: ____ Current Dwelling Amount: ____
Policy Start Date ____ Known Losses (past 5 years): ____

Home & Property Information:

Physical Address: ____ Year Constructed: ____ # Living in Home: ____ # of Families: ____
____
Miles from Fire Dept. ____ Responding Fire Dept. ____
Fire Hydrant Near: ____ Inside City Limits? ○ Yes ○ No
Wood Stove: ○ Yes ○ No Fireplace: - Fireplace - Age of Roof (Years): ____ Roof Type: ____
Above Ground Sq. Footage: ____ House Style: - House Type - % of Basement Finished: ____
Walkout Basement: ○ Yes ○ No Age of Plumbing (Years): ____ # Full Bath: ____ # ¾ Bath: ____ # ½ Bath: ____
Heat Source & Type: ____ Age of Heat Source (Years): ____ A/C Central Air: ○ Yes ○ No
Age of Electrical (Years): ____ Kitchen Type: - Kitchen Type - Specialized Doors: ____
Other Special Features: ____ Business Usage: ○ Yes ○ No
Garage: - Garage Type - Garage Size: - Garage Size -
Attached Structures: ____ Size of Structure: ____ Security System: ○ Yes ○ No
Attached Structures: ____ Size of Structure: ____ Total Acreage: ____
Other Attached Structures: ____ Size of Structure: ____ Trampoline: ____
Swimming Pool: ○ Yes ○ No Pool Type: - Pool Type - Pool - Pool Extras -

Phone: 763-682-4890
Fax: 763-684-5278
Email: info@insurancecenterofbuffalo.com
Address: PO Box 458, Buffalo MN 55313

Y = Sale prices of homes

Exploring the dataset

In order to drastically reduce the number of features, we need to find:

- Strong predictors (to keep)
- Skewed values (to drop)
- Null values (to fill up)
- Repeated/ similar questions that can be combined (to reduce features)
- Change as many categorical questions into binary/ ordinal questions
- Outliers (to drop)

Strong predictors against the sale price

9 Features	Correlation against sale price
Overall quality	0.800
Gross living area	0.697
Garage area	0.650
Cars that can fit in garage	0.648
Total basement area (sqft)	0.629
1st floor area (sqft)	0.618
Year house was built	0.572
Year remodelled	0.550
No of full baths	0.538

- Some of the important predictors include the quality, area and age features.

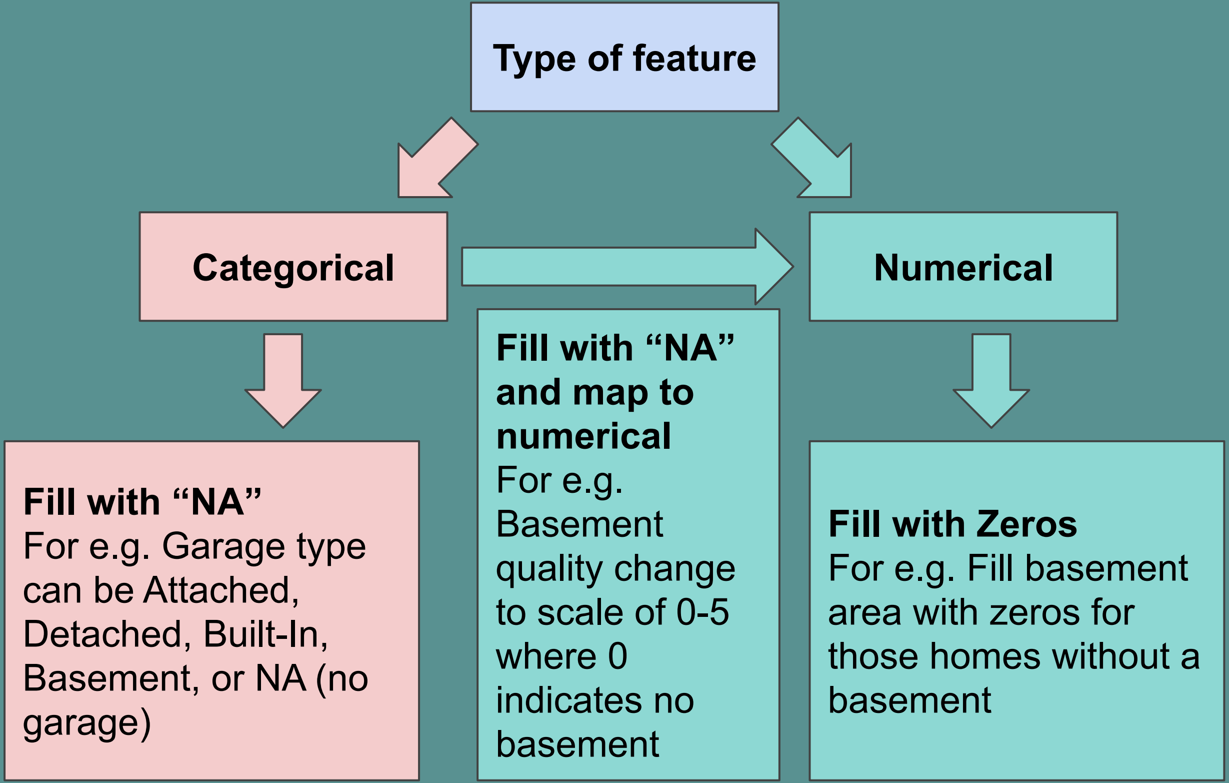
- We will keep and use these predictors as questions in our application form.

Skewed features

Features	Highest % of a single value	Interpretation
Alley	93.17	93% of homes do not have an alley
Utilities	99.90	99% of homes have electricity, gas, water and sewage
Land slope	95.22	95% of homes are built on a gentle slope
Roof material	98.73	99% of homes have roof made of composite shingles

- Some features have a very high % of a single value. Such features are usually poor predictors and can be dropped.
- For e.g. Asking a customer if his house has utilities in this day and age is seemingly redundant.
- **Recommendation: Drop most of the skewed features**

Null values



Type of feature

Categorical

Numerical

Fill with "NA"
For e.g. Garage type can be Attached, Detached, Built-In, Basement, or NA (no garage)

Fill with "NA" and map to numerical
For e.g. Basement quality change to scale of 0-5 where 0 indicates no basement

Fill with Zeros
For e.g. Fill basement area with zeros for those homes without a basement

We expected null values in the dataset for the following reasons:

- Clerical mistakes
- Incomplete form
- No suitable values

Recommendation:

- **Fill up categorical questions with "NA". Convert any possible categorical features into numerical ones.**
- **Fill up numerical questions with zeros**

Prevention Measure

- **The new form will have data validation and restrict free text.**

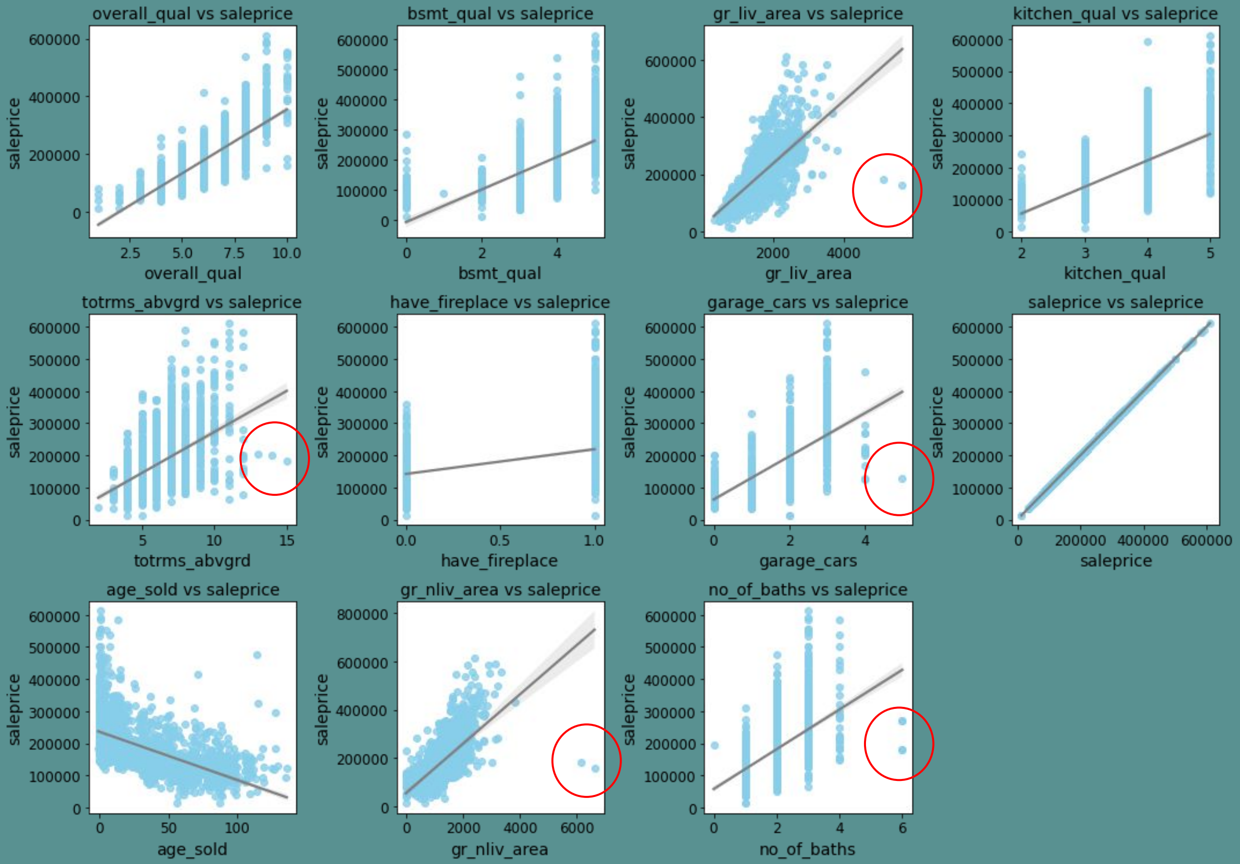
Combining redundant questions

Feature 1	Feature 2	Correlation	Interpretation
Year garage was built	Year house was built	**0.83**	Garage & House built together
Garage area	Cars in garage	**0.89**	Bigger the garage, the more cars it can fit
First floor area	Basement area	**0.81**	Basement area likely to increase in tandem with first floor area

Multicollinearity: This is when an independent variable is highly correlated with another independent variable. We have to get rid of it to have more reliable inferences.

Recommendation: Keep only one independent term and drop other highly correlated terms

Outliers



Contributes to large errors during regression analysis

Use a scatter plot to quickly identify and remove outliers

Recommendation: Drop to get a better line fit for each independent variable

After preprocessing

Categorical features

Neighbourhood
MS Subclass
MS Zoning
House style
Exterior
Masonry veneer type
Foundation

Numerical features (with correlation to sale price)

Gross living area (0.72)
Gross non-living area (0.70)
Cars in garage (0.66)
Year built (-0.57)
No of baths (0.60)
Fireplaces (0.49)
Total rooms (0.51)
Overall Quality (0.80)
Basement Quality (0.61)
Kitchen Quality (0.70)

We are left with 7 categorical features and 10 numerical features.

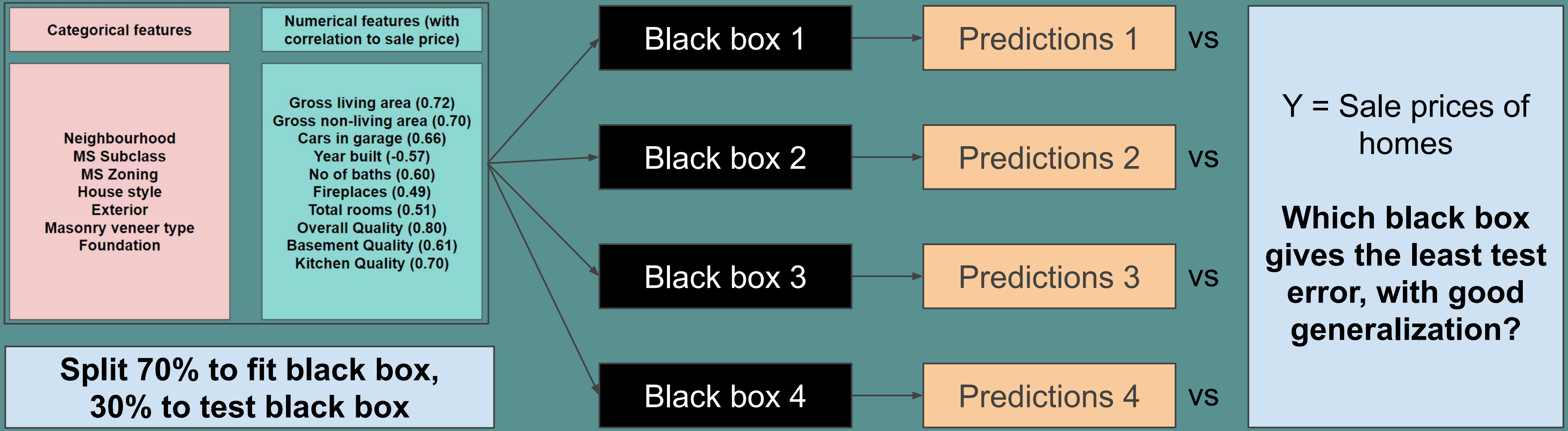
Numerical features show a high correlation to the sale price, which should give us a good model prediction.

Black box to test our models

We throw our features into 4 different black boxes to generate a predicted sales price.

Each black box contains different parameters to optimise the model score.

We want to find the black box that gives us the least test error with good generalization.



Categorical features

Neighbourhood
MS Subclass
MS Zoning
House style
Exterior
Masonry veneer type
Foundation

Numerical features (with correlation to sale price)

Gross living area (0.72)
Gross non-living area (0.70)
Cars in garage (0.66)
Year built (-0.57)
No of baths (0.60)
Fireplaces (0.49)
Total rooms (0.51)
Overall Quality (0.80)
Basement Quality (0.61)
Kitchen Quality (0.70)

Split 70% to fit black box, 30% to test black box

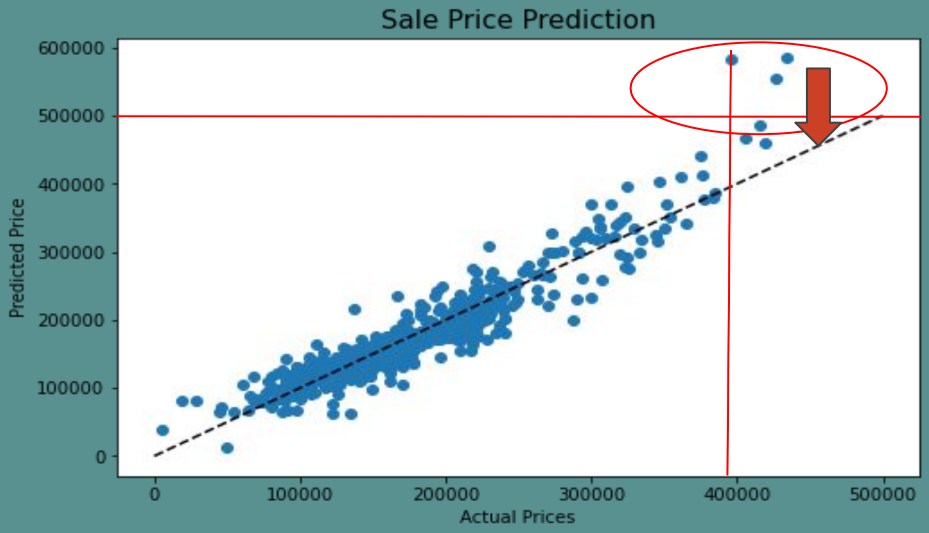
Y = Sale prices of homes

Which black box gives the least test error, with good generalization?

Results

- Black box 2 performed the best with a 1.97% difference between the train and test set (good generalization) and also has the small test error (low variance)
- Predicted and actual values fit nicely on a linear graph
- However model does not perform well when predicting above $390k
- We can improve on model by offsetting all predicted values that are $500k and above (so that it moves closer to the dotted line)

Model	Error (train)	Error (test)	Difference (%)
Black box 1	26687	27256	-2.13%
Black box 2	26696	27222	-1.97%
Black box 3	26690	27236	-2.05%
Black box 4	29482	29577	-0.32%



Final application form

Please choose your neighbourhood, MS Subclass and MS Zoning (these are found in your house document sheet)
Please fill in your house style (1 story or 2 story or Others)

How big is your living area? (add first floor and 2nd floor sqft)
How big is your non-living area? (basement area and porch area sqft)

How many rooms are there?
How many full baths are there?
Do you have a fireplace?
How many cars can your garage fit?

When was your house built? (enter year)
When was the application form filled up (Can be calculated back end = Mths Sold)

What is the exterior finishing on your house
What is your masonry veneer type? (Brick, stone, others or none)
What kind of foundation is your house on? (Cinderblock, Concrete, Bricktile or others)

What is the overall material and finish quality of your house (1 to 10)
What is the height of your basement? (convert 1 to 5)
What is your kitchen quality? (1 to 5)

Please click here to fill up the form.

Conclusion

We were able to get a sufficiently good model that is relevant for our use case:

- **Shortens drastically the time** a customer takes to fill the form
- Decreases the **dropout rate** and **increase revenue**
- **Simple and easy to understand** application form
- **Scalable** savings

We can improve the process further by:

- Create and deploy a simple web application for customers to get instant results instead of through email
- For higher valued homes:
    - Model may not perform as well when prices exceed roughly USD 390,000 and up
    - It would be advisable that an on-site valuation be handled by human intervention
- Alternatively, gathering of data for the higher valued homes will also help in assisting the model to better predict home prices in excess of USD 390,000
- Get house pricing data from other years instead of during the Financial Crisis

APPENDIX

Size

Gross living area (gr_liv_area)
Gross non-living area (bsmt area + porch area = gr_nliv_area)
GarageCars: Size of garage in car capacity

Profile

Neighbourhood
MS subclass
MS zoning
House style (1 or 2 story or others)

Time factors

Year built

Features

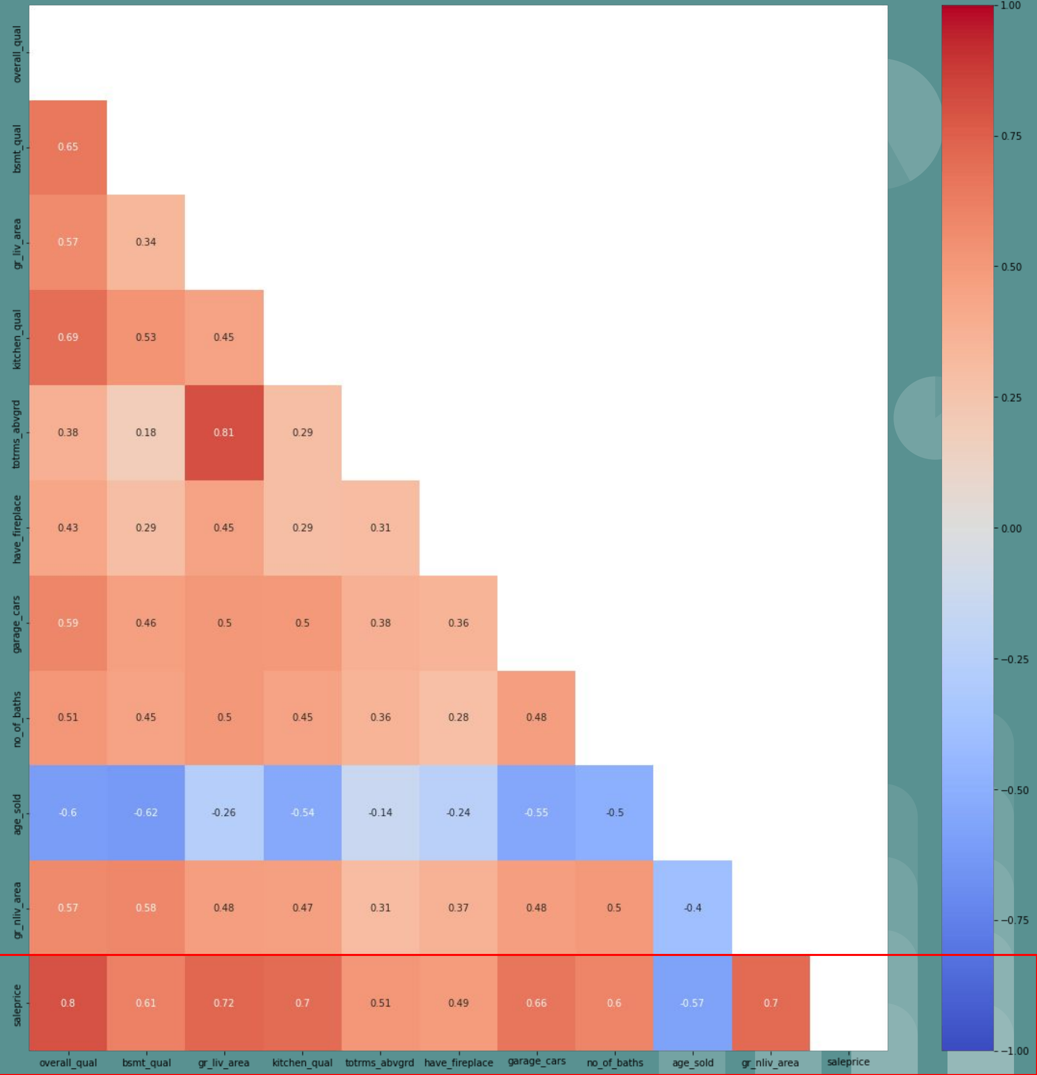
No. of baths
Fireplaces: Number of fireplaces
Total rooms above grade

Finishing

Exterior1st: Exterior covering on house
Masonry veneer type (Brick, stone, others or none)
Foundation (Cinderblock, Concrete, Bricktile or others)

Quality

OverallQual: Overall material and finish quality
BsmtQual: Height of the basement
KitchenQual: Kitchen quality


overall_qual
bsmt_qual 0.65
gr_liv_area 0.57 0.34
kitchen_qual 0.69 0.53 0.45
totrms_abvgrd 0.38 0.18 0.81 0.29
have_fireplace 0.43 0.29 0.45 0.29 0.31
garage_cars 0.59 0.46 0.5 0.5 0.38 0.36
no_of_baths 0.51 0.45 0.5 0.45 0.36 0.28 0.48
age_sold -0.6 -0.62 -0.26 -0.54 -0.14 -0.24 -0.55 -0.5
gr_nliv_area 0.57 0.58 0.48 0.47 0.31 0.37 0.48 0.5 -0.4
saleprice 0.8 0.61 0.72 0.7 0.51 0.49 0.66 0.6 -0.57 0.7
overall_qual
bsmt_qual
gr_liv_area
kitchen_qual
totrms_abvgrd
have_fireplace
garage_cars
no_of_baths
age_sold
gr_nliv_area
saleprice