BROWN
DATA
SCIENCE

Brown Datathon 2020:

David Fan | Kai Lu | Yi Wang | Wanxin Ye | Weihao Zhou

Procedure and Approach

Our Goal is to make prediction despite the noise and randomness - the overarching strategy is foundational in its success

On a zip9 level, the rampant existence of noises is likely going to overshadow the true generalizable trend

In more detail:

External Data
Macroeconomic data on **zip5** level

Merge
With demographic data aggregated to **zip5** level

Classification
Classify (0, 1) active purchase zones and inactive purchase zones

Regression
For the active zones (1), predict a continuous count for each zip5

Citizen bank's Data
On **zip9** level

Merge
demographic & all finance data on **zip9** level

Regression
Use **proportion of zip9 homebuyers over zip5 level** as label to train the entire dataset

Allocation
Zip 5 prediction **x zip 9** level proportion to get final prediction

Picking the right model to account for non-linear features is crucial to the accuracy of the prediction

At all stages (Classification, Regression & Allocation), we employed three main models each with proper hyperparameter tuning:

Baseline

Simple Regressions

We used the Logistic Regression **(Classification Stage)** and Linear Regression **(Regression Stage)** combination to create a baseline accuracy to improve upon

Ensembles

XGboost

RandomForest

XGBoost and Random Forest were then used in both classification and regression as an attempt to further optimize the results with **both bias and variance reduction along with accomodation for non-linearity.**

Feature Selection

Our feature is our story - every aggregated observable is prompted by macroeconomic element that affects the underlying propensity to buy

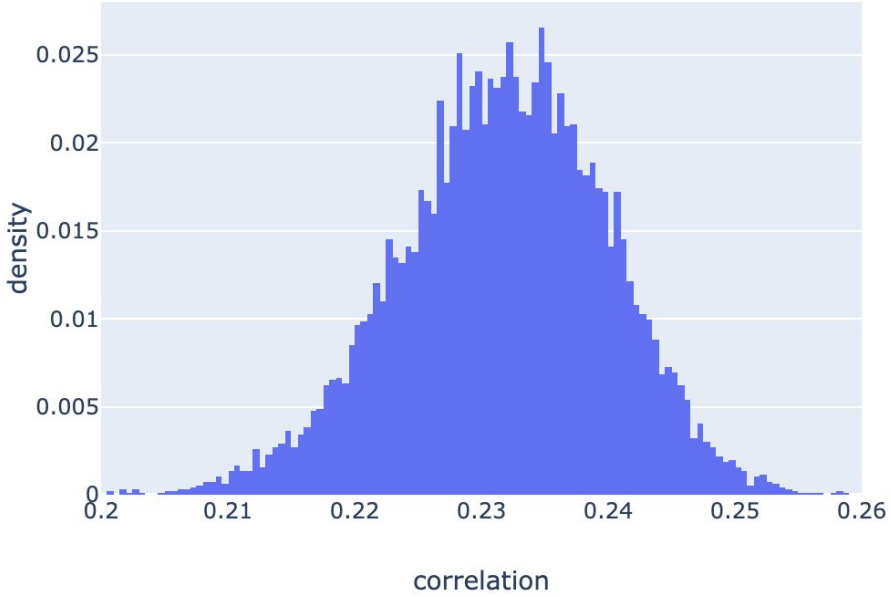
Due to the income effect, households in less affordable areas had higher propensity of buying homes

Rationale: **Economic Advantage**

In the right plot we **bootstrapped the correlation between affordability** (measures the years of working needed to afford median house price) and **number of home buyers in each zip code.** There is a noticeable relationship between the two which suggests that as houses become more unaffordable, the more likely we are to see home buying behavior.

Features: **Median Salary** (IRS) **& Median House Prices** (Zillow)

Correlation of **Affordability** vs **# of Buyers**



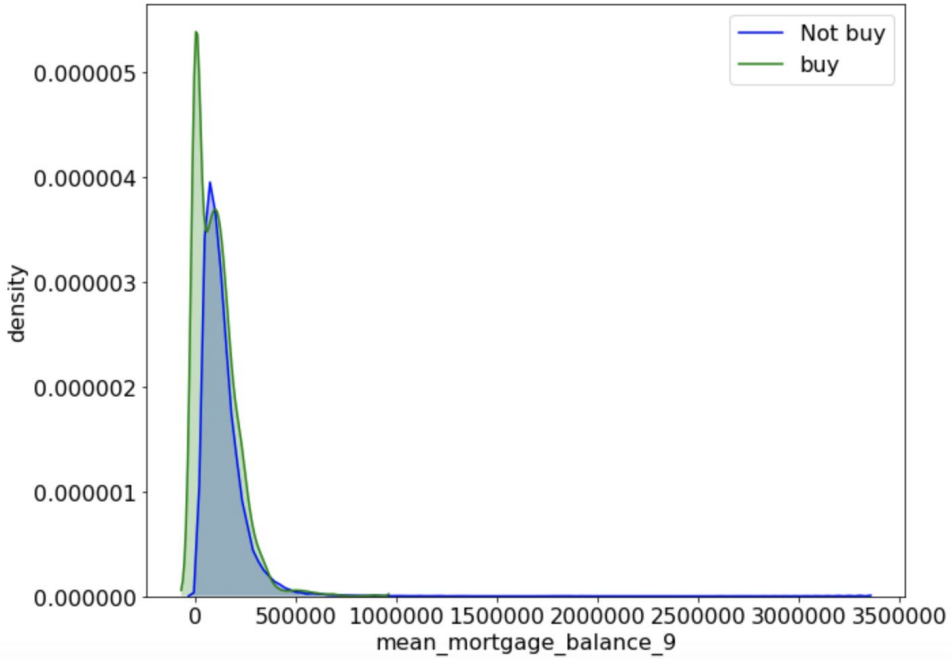
Group by homebuyers indicator, households with less involving debts had higher propensity of buying homes

Rationale: **Economic Advantage**

In the right plot we have two conditional class distribution: mean_mortgage_balance_9 given homebuyers > 0 in corresponding zip 5 and mean_mortgage_balance_9 given homebuyers = 0 in corresponding zip 5. From the graph we can tell that **positive class distribution is more peaked and skewed to the right**, **implies the lower mean mortgage balance in a zip 5 area, the higher likelihood people from that area will purchase new home.**

Features: **Mean Mortgage Balance & Home Purchase Indicator**

Distribution of **Mean Mortgage Balance** given **Home Purchase Indicator**



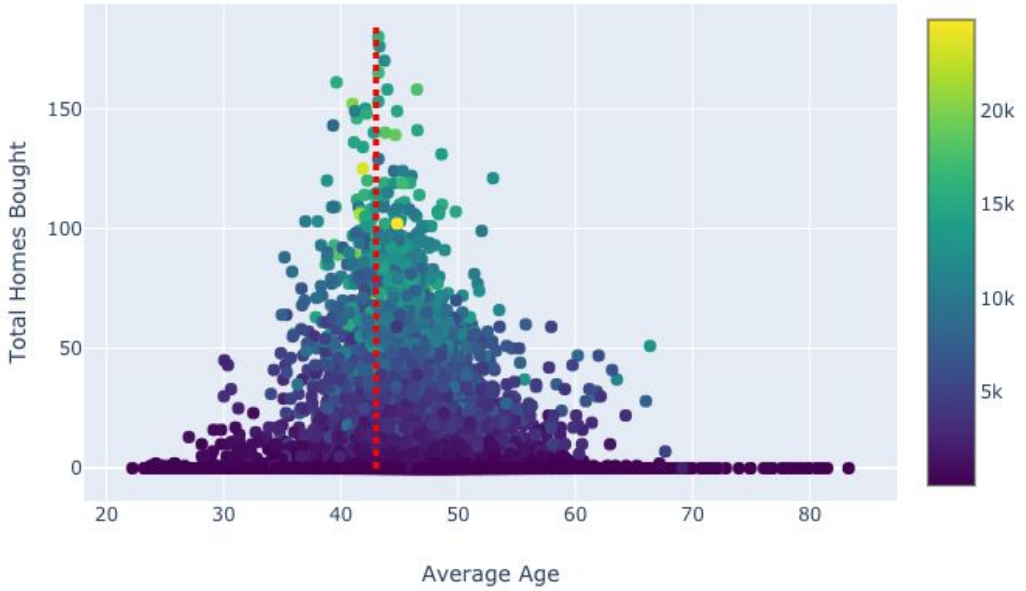
At a certain age in life, people arrive at the point where they want to buy a house. After that age, the propensity drops rapidly.

Rationale: **Maturity**

The plot on the right shows that **after a certain point in age, propensity of home buying decreases.** This point is most notable at age 42. The color bar shows the number of households in each zip code. Naturally, there is a multiplicative effect with more households leading to higher amounts of home purchasing in those zip codes.

Feature: **Average Age**

Scatter of **Total Home Brought** vs **Avg Age**



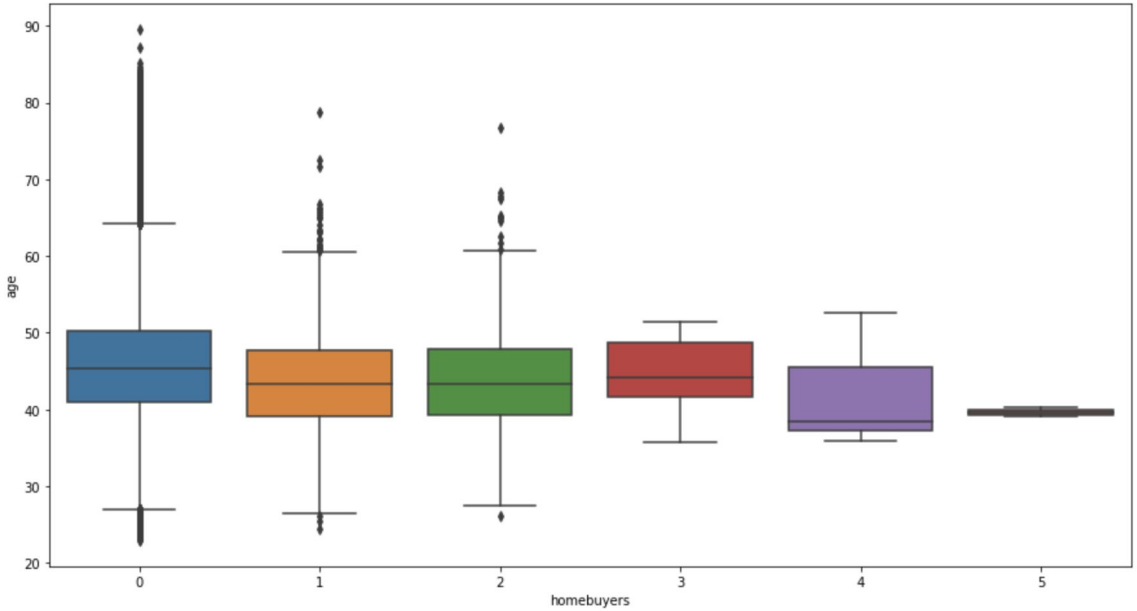
At a certain age in life, people arrive at the point where they want to buy a house. After that age, the propensity drops rapidly.

Rationale: **Maturity**

The plot on the right shows that **number of home buyers in each zip9 area is sort of negatively correlated with the avg age in that area.** This is consistent with previous finding, we have the highest avg age in the zero home buyers sub group and the lowest avg age in the four/five home buyers sub group.

Feature: **Number of Home Buyers in Each Zip9**

Box Plot of **Avg Age** vs **Number of Home Buyers in Each Zip9**



Propensity of buying is additive. The more households in a zip code, the underlying rate of buying increases due to more opportunities.

Rationale: **Multiplicity**

Each **individual** has some unobservable propensity p of buying a home. **Multiple individuals** have an aggregated higher propensity to buy.

It is necessary to control for the number of individual when making predictions to account for the **difference in opportunities (heterogeneity)** across different zip-codes.

Feature: **# of total household in the zip-code**

Results and Interpretation

After aggregating to zip5 level, we measure our performance at each step of the process: running a classification model, regressing on non-zero predicted zip5s, and finally allocating down to zip9s grain for our final predictions.

A sequential model implies the necessity to evaluate the results at each individual process

We used a 80% train-20% test split, and picked the best performing model at each stage:

1 Our best performing clustering algorithm, by accuracy, was the **XGBoost**, with the confusion matrix attached below

2 The best performing regression algorithm overall was the **Random Forest**, we included the test accuracies for both the non-zero test data points and the overall (0 concatenated test data points)

		Actual	
		Positive	Negative
Predicted	Positive	145	66
	Negative	132	2307

Regression Results

	Non-zero RMSE	Non-zero MAE	First Timer RMSE	First Timer MAE	Overall RMSE	Overall MAE	Weighted Loss
RF	11.1	7.4	6.55	4.2	9.75	6.15	611.54
Mean	21.7	16.4	11.6	8.6	20.2	14.9	3144.995

A sequential model implies the necessity to evaluate the results at each individual process

We used a train - test split, and picked the best performing model at each stage:

3

The allocation model is trained on the zip9-level covariates to predict its expected proportion with respect to its zip5 counterpart; the best model as seen below is the XGBoost

Allocation Model Results

	First Timer RMSE	First Time MAE	Overall RMSE	Overall MAE	Loss Weighted
RF	0.0591	0.0091	0.0505	0.0070	0.0162
XGB	0.0052	0.0073	0.0430	0.0052	0.0161

A sequential model implies the necessity to evaluate the results at each individual process

We used a train - test split, and picked the best performing model at each stage:

4

We will now put everything together, to make our final prediction on the zip9 level. This will simply be the expected buyer count at a zip 5 level multiplied by the expected proportion at a zip 9 level times a scaling factor.

Final Sequential Model Results

	First Timer RMSE	First Time MAE	Overall RMSE	Overall MAE	Loss Weighted
0	0.179	0.023	0.023	0.04	0.60
Fin	0.11	0.012	0.020	0.027	0.121

Thank you!

Any Questions?