Lending Opportunity

Bic Vu

AI Guild Apprentice Program Capstone

Cohort 11 June 2023

Can you buy into better opportunities?

Social capital — “the strength of our relationships and communities” — has been identified as a determinant of upward mobility. A study on social captital[1] showed that “living in a place that fosters these connections causes better economic outcomes.”

When people are looking to buy a home, especially as a family, they consider the opportunities of that location. But what if the existing social capital network of that location has an impact on the likelihood of their home loan getting approved? How does the existing economic connectedness, cohesiveness and civic engagement of a location impact a potential home buyer’s ability to access opportunities?

Proposed Solution

Analyze the potential correlations between Social capital features along with standard economic inputs from CFPB’s Home Mortgage Disclosure Act (HMDA) data for the location of the property.

Build and test models that predict the probability of a home approval.

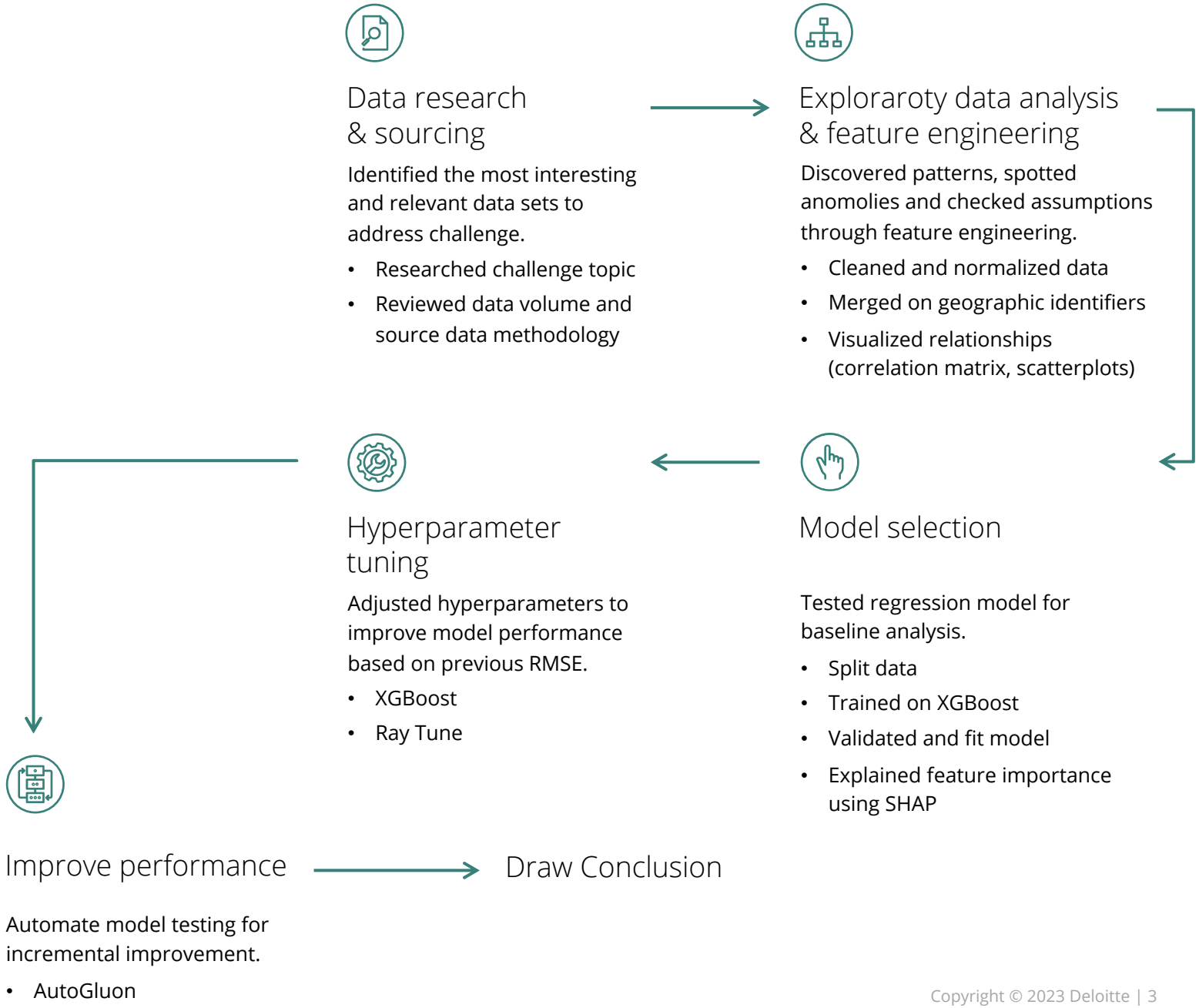
Impact

Beyond standard characteristics of a neighborhood such as schools and amenities that makes a home appealing, potential lenders could also consider the location of the home as an opportunity to access a better social capital network.

1. Chetty, Jackson, Kuchler, Stroebel, et al. *Nature* 2022

Analytics Pathway

Process from initial research, exploratory data analysis to model development and tuning.



ETL & EDA

About the Data

Research and exploratory data analyses.

What makes this data interesting? People!

The Consumer Financial Protection Bureau's (CFPB) Home Mortgage Disclosure Act (HMDA) data contains financial statistics about people, such as debt to income ratio, that directly impacts a loan appraisal, along with potentially influencing statistics such as race and percentage of minority population in the census tract.

The Social Capital data measures economics opportunities of by using Facebook data to determine metrics such as the ratio of high and low income friends in a community.

By linking the location of a home loan with the economic opportunity of that geographic area, the study aims to observe correlations between a community's social connectivity and opportunity to buy into that community.

HMDA

Economic metrics

- Loan approval
- Income
- Debt to income
- Loan amount
- Property value

Socio-geographic metrics

- Race
- Minority population
- Census tract income level

Merged on Census tracts

Converted zip to Census tracts

Social Capital

Economic metrics

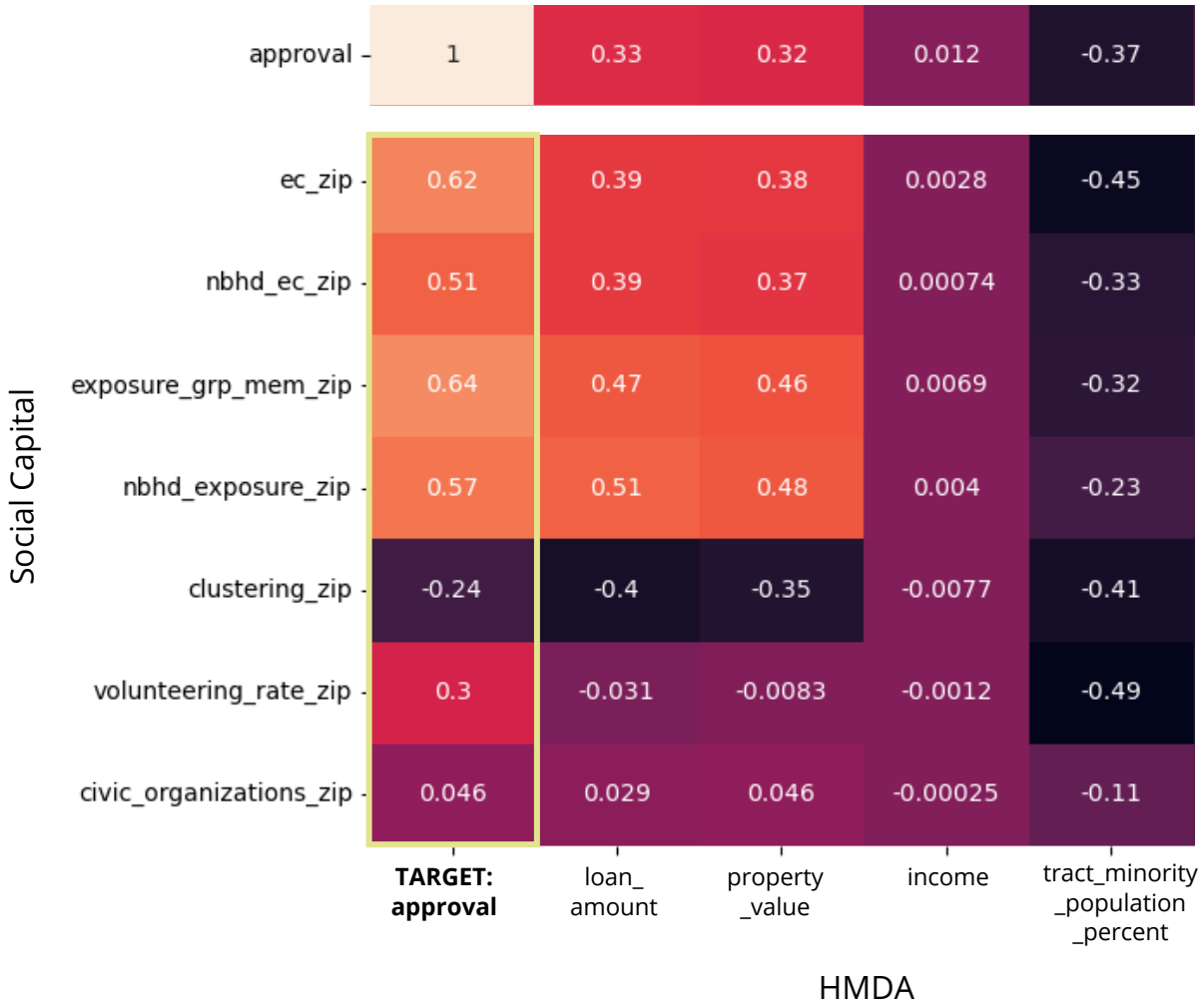
- Economic connectedness
- Exposure to high income individuals

Socio-geographic metrics

- Clustering of friends
- Volunteering rate
- Civic organizations

Reference data dictionary in appendix section

Features Correlation



Analyzing relationships

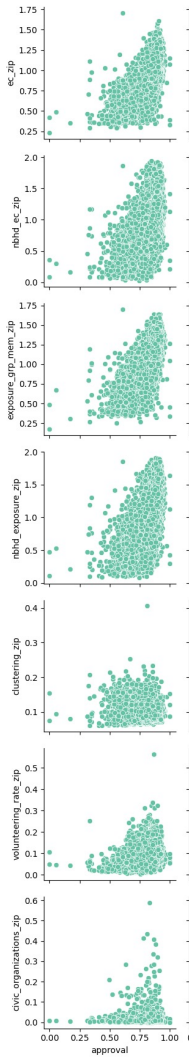
The correlation matrix reveals that there are highly positive linkage between economic connectedness and exposure metrics in the Social Capital data with the approval rate for home loans in the HMDA data.

Secondary high correlations exists between exposure and loan amount and property value

Other relationship were minimal or negative. Particularly negative was the low level economic connectedness with tracts where a large percentage of the population is minority.

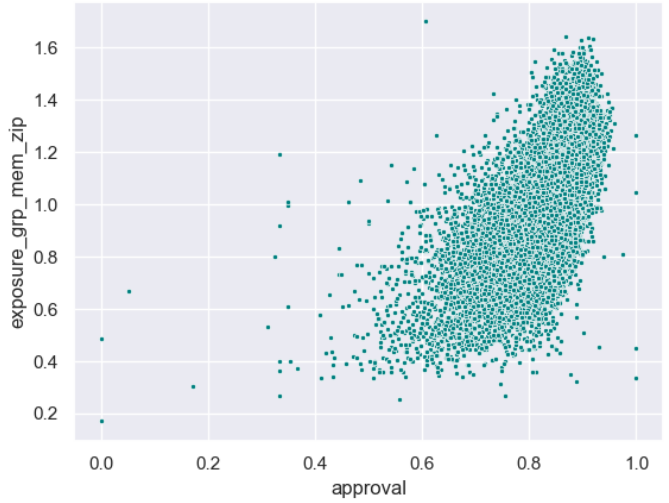
Similarly negative is the correlation between voluntering rate and minority population, indicating that there is perhaps a low rate of volunteering in Census tracts where there is a high minority population.

Scatter Plot of Key Features

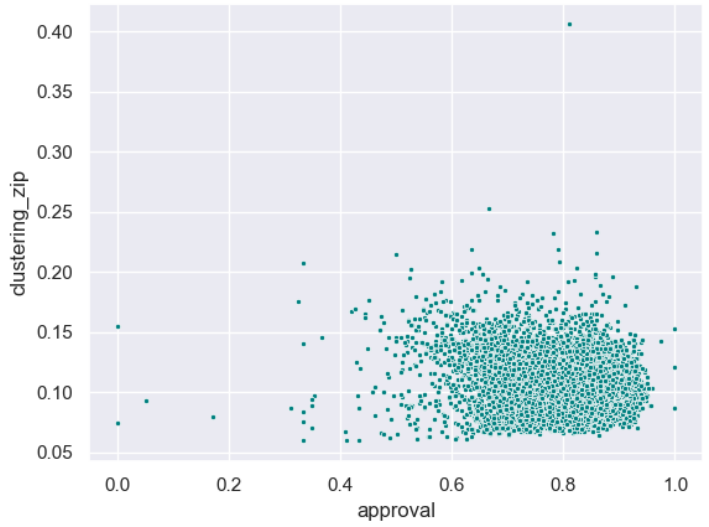


Feature sample (Social Capital)

Target (loan approval)



Scatterplot of highest positive correlation feature (exposure to high income friends) with target (loan approval)



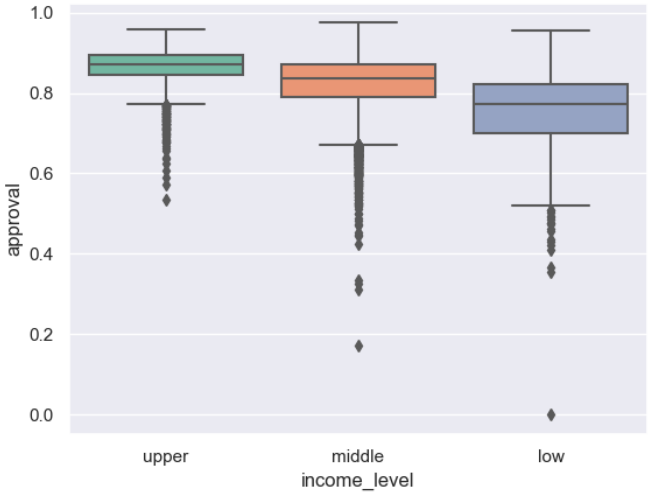
Scatterplot of lowest negative correlation feature (clustering) with target (loan approval)

Distribution of Correlations

Scatterplots were used to visually observe patterns in relationships of features with highest and lowest correlations compared to the target.

Additionally box splots were used to reference check categorical features, such as income level of Census tracts.

The various non-linear patterns of distribution shown in the scatterplots comparing target to features makes this data set a candidate for a regression model.



Box plot of income level for Census tracts compared to loan approval

MODEL 1

XGBoost Regressor

Model Selection

The EDA process indicated that there are non-linear patterns in the data which makes this challenge a candidate for a regression model. (The target metric of loan approval is also a continuous probability ranging 0 to 1, which rules out classification.)

An XGBoost regressor was selected for the initial model because of options for early stopping to avoid overfit. Results were fairly accurate, particularly predictions for approal rates above 80%.

Model Type	XGBoost Regressor
Key model adjustments	n_estimator = 300 max_depth = 6 early_stopping_rounds = 10
Best validation RMSE	0.04720

Prediction accuracy (RMSE)

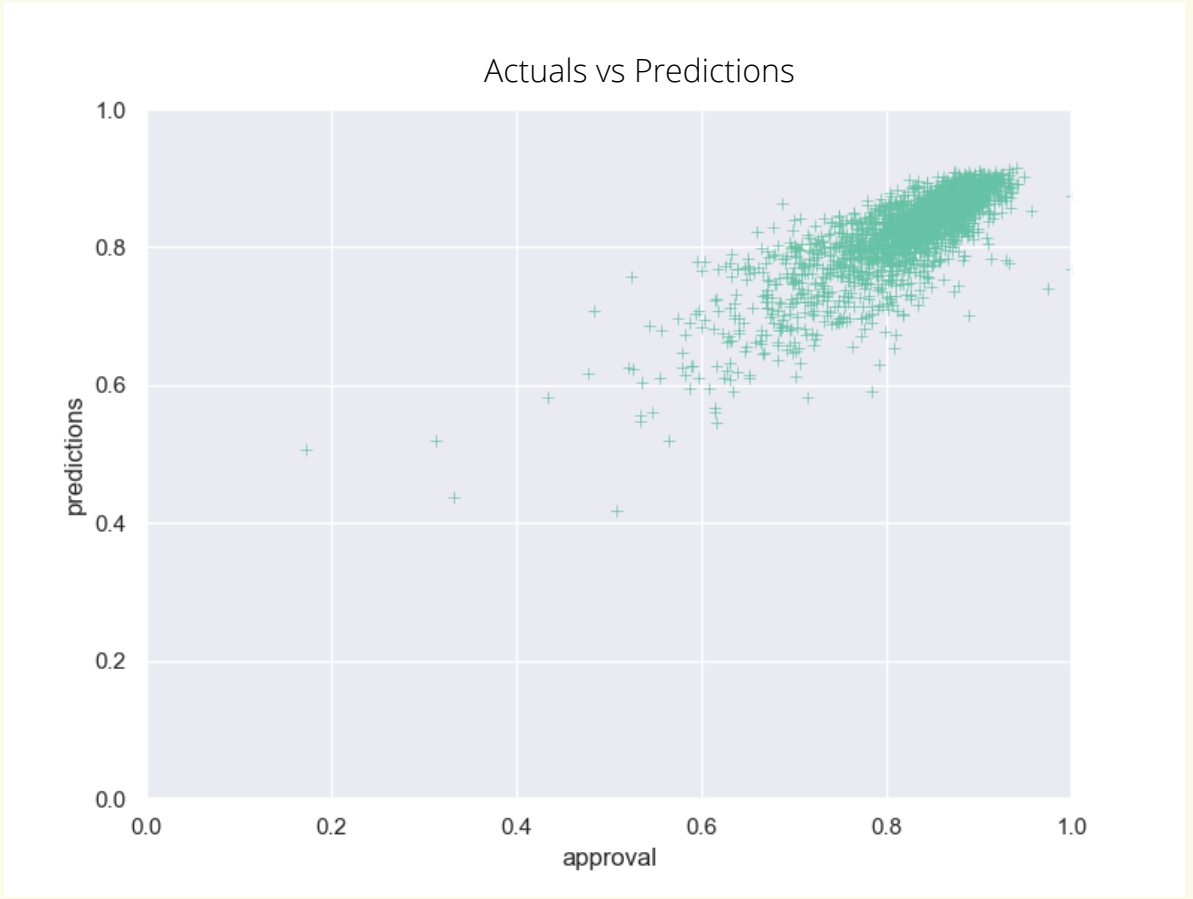
Model 1
XGBoost
0.044423

Target Metric
approval[1]

1. *Percent of home loans approved in a census tract out of all approved or denied applications*

Data split for all models

Train	Validate	Test
70% 9,979 rows (one Census tract per row)	15% 2,139 rows	15% 2,138 rows



Scatterplot comparing actual targets for approvals (x axis) and predictions (y axis) from XGBoost Regressor model.

MODEL 2

Hyperparameter Tuning: XGBoost with Ray Tune

Improve performance with hyperparameter tuning

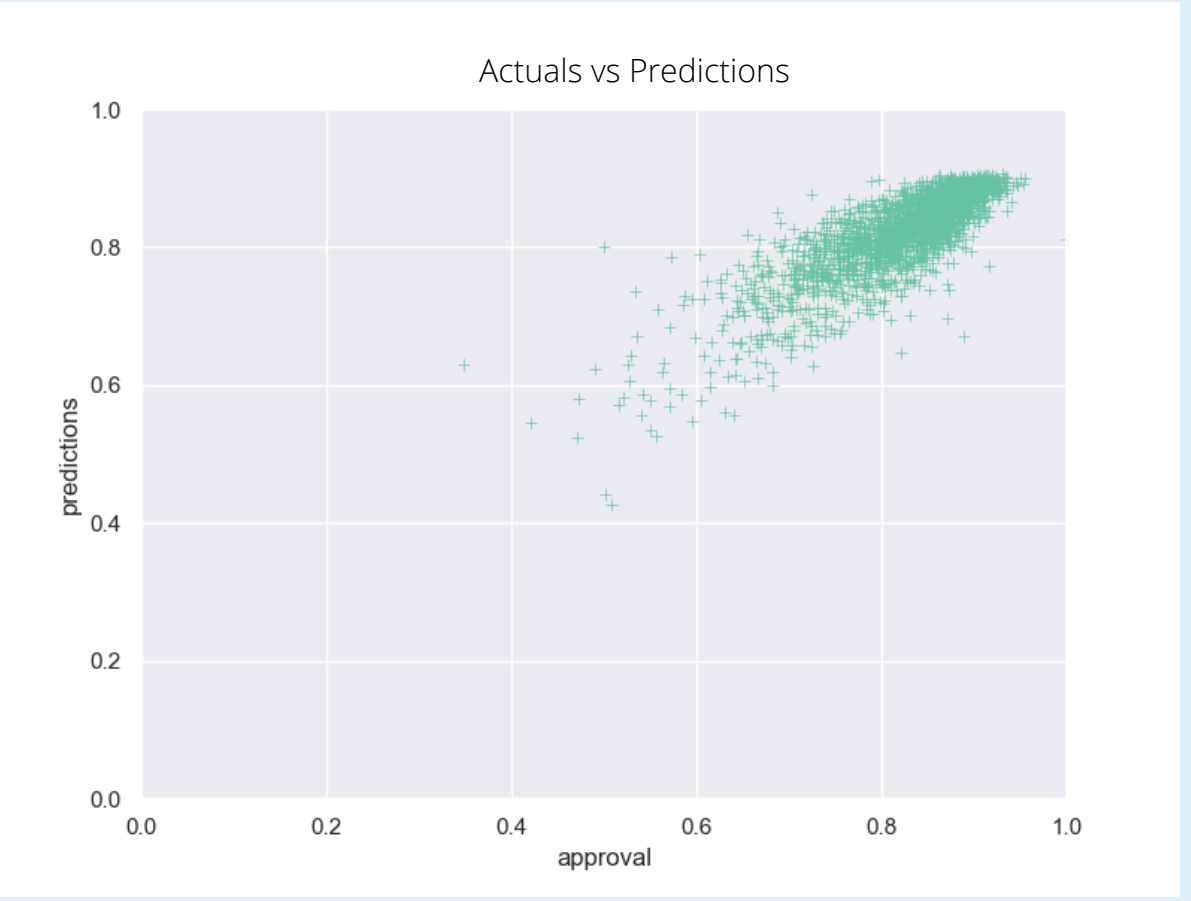
To improve the accuracy of the model and hone in on an optimal set of hypeparameters to tune, Ray Tune was used to run 100 training tests and return the best peformaning model. Based on the model with the best validation RMSE, a new XGBoost model was developed and used for predictions.

Model Type	XGBoost Regressor
Key model adjustments from 100 Ray Tune trials	n_estimator = 300 max_depth = 6 early_stopping_rounds = 10
Best validation RMSE	0.045859

Prediction accuracy (RMSE)

Model 1
XGBoost
0.044423

Model 2
XGBoost & Ray Tune
0.043132



Scatterplot comparing actual targets for approvals (x axis) and predictions (y axis) from XGBoost Regressor model adjusted with Ray Tune output.

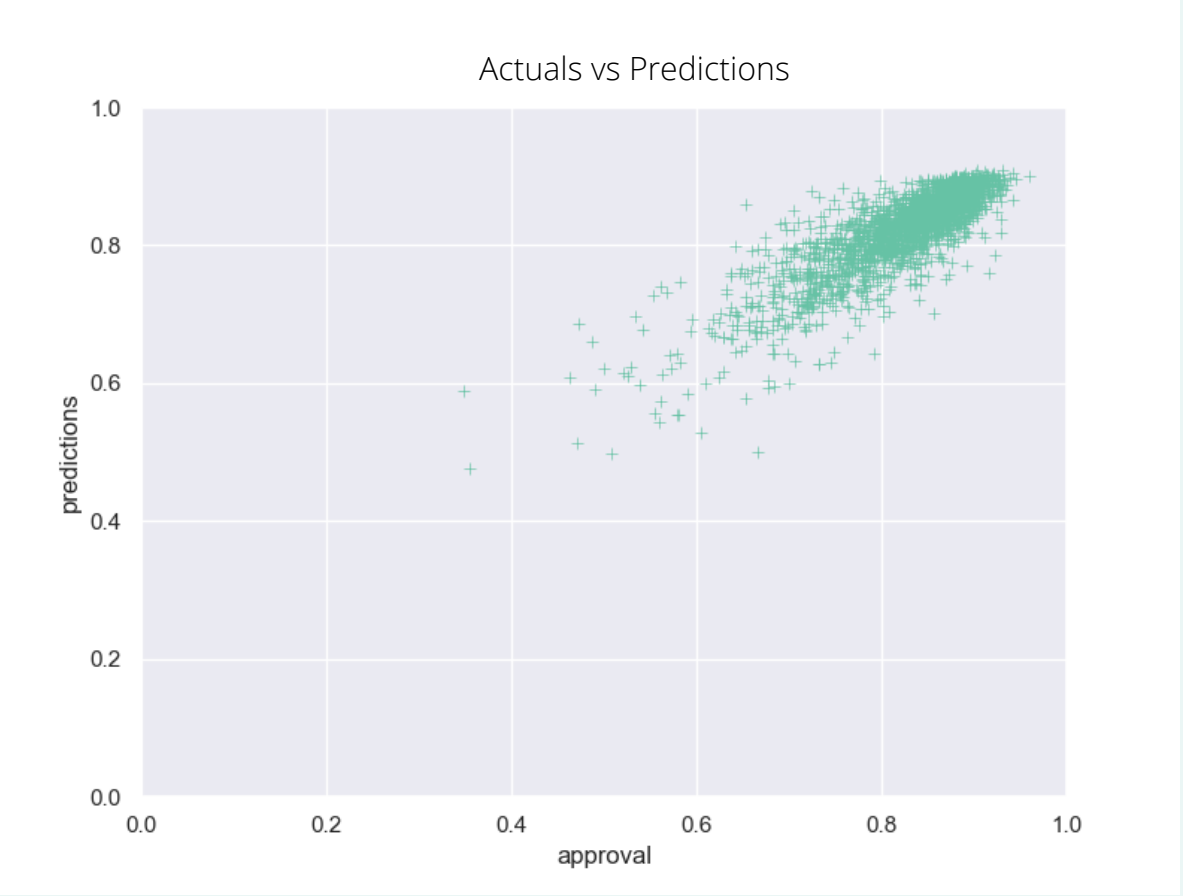
MODEL 3

AutoGluon

Improve performance with different model types

While Ray Tune optimized the performance of XGBoost, there are additional opportunities to increase performance by trying different model types. AutoGluon allowed for a run on 14 different model types and return the best model based the given RMSE evaluation metric.

Model Type	WeightedEnsemble_L3
Key model adjustments	Changed model type to WeightedEnsemble_L3 based on RMSE
Best validation RMSE	0.039855



Scatterplot comparing actual targets for approvals (x axis) and predictions (y axis) from Weighted Ensemble L3 model.

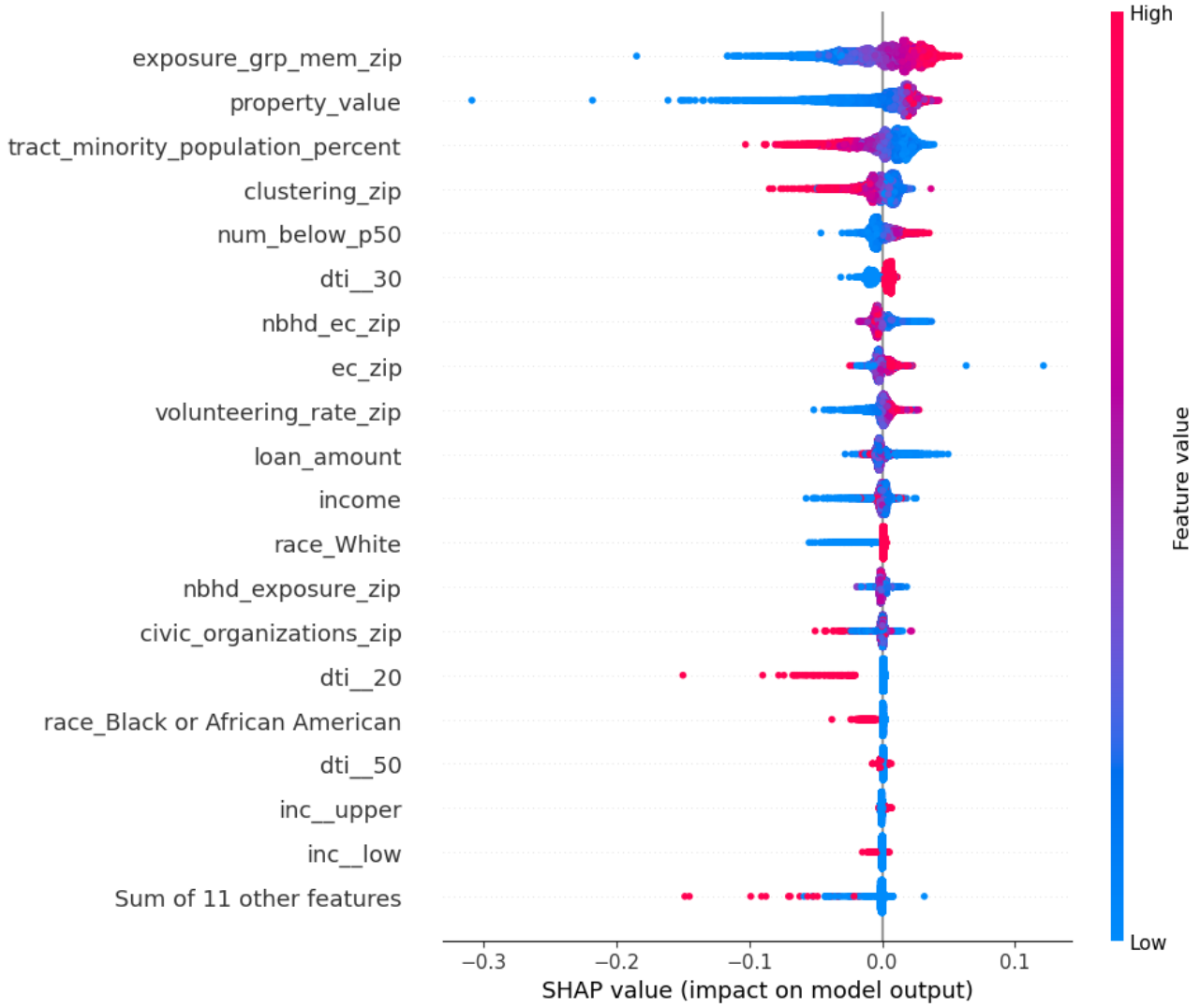
Prediction accuracy (RMSE)

Model 1
XGBoost
0.044423

Model 2
XGBoost & Ray Tune
0.043132

Model 3
WeightedEnsemble_L3
0.039900

Model Explanation



Feature Importance Evaluation

The SHAP explainer was used to review the impact of features on the output of the two XGBoost Regressor models (model 1 & model 2). A number of the Social Capital data ranked highly in feature importance, meaning they had a high impact on the predictions.

Exposure to high income friends (exposure_grp_mem_zip) in particular had the highest positive impact on predictions. That is, when a loan application was for a property in a Census tract where residents were exposed to more high income friends, the loan was more likely to be approved.

Economic exposure (ec_zip) had a fairly high impact on a concentrated number of predictions. Similar to the correlation matrix, clustering had a negative correlation.

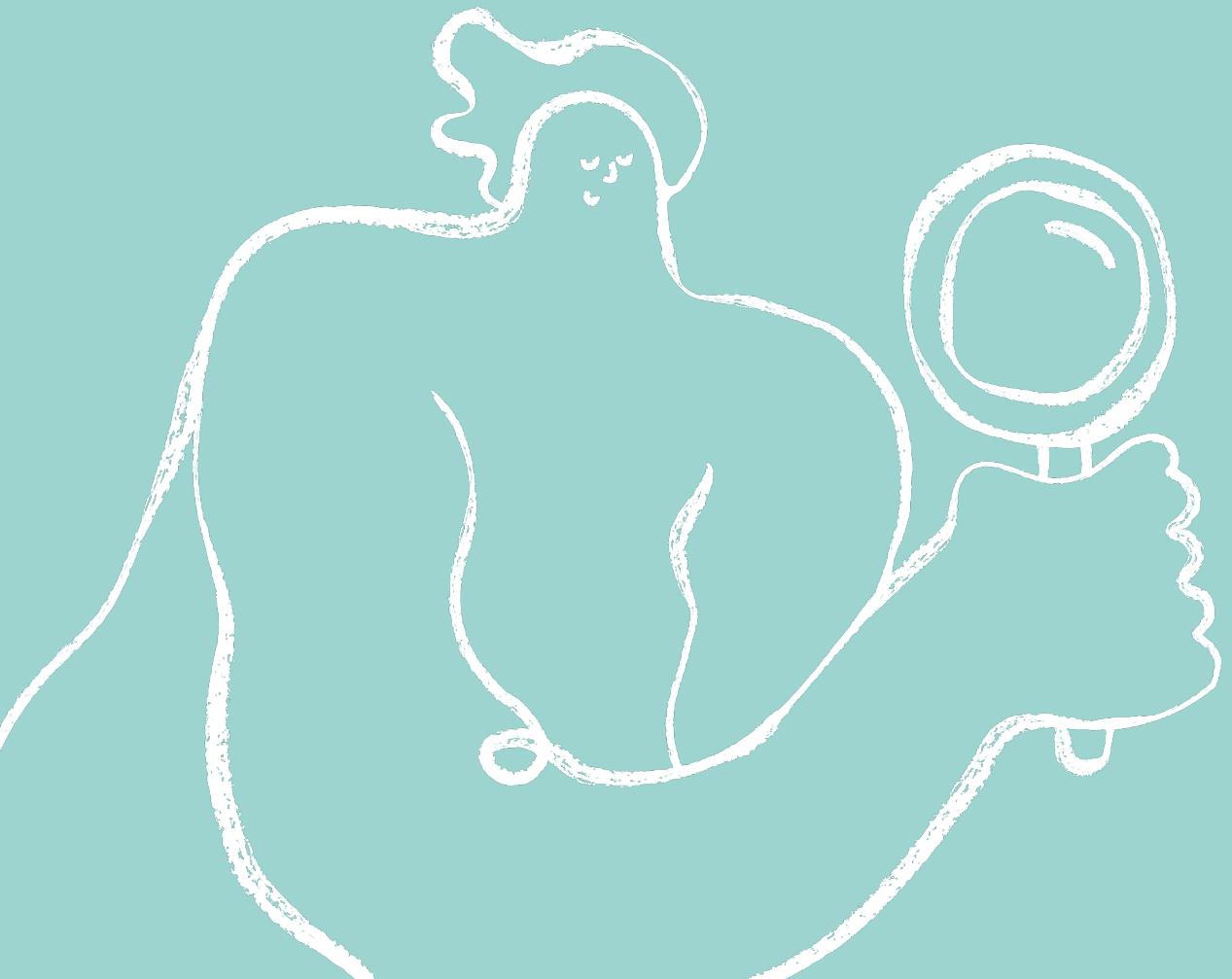
For a small number of tracts where applicants were predominantly white, it had a concentrated positive impact. The opposite is true of Black applicant with a small concentration of negative impact.

SHAP beeswarm of feature importance from highest to lowest

Conclusion

We might be peripherally aware that our social network has an impact on future opportunites, but through this study, we can see that the existing social network of a home's location can impact a lender's chance of getting a home loan approved.

Extending on Opportunity Insight study, this analysis shows that economic connectedness and exposure to high income friends not only reinforces the social network patterns in place but can also be a barrier to entry for someone looking to buy into that community.



Learning Outcomes

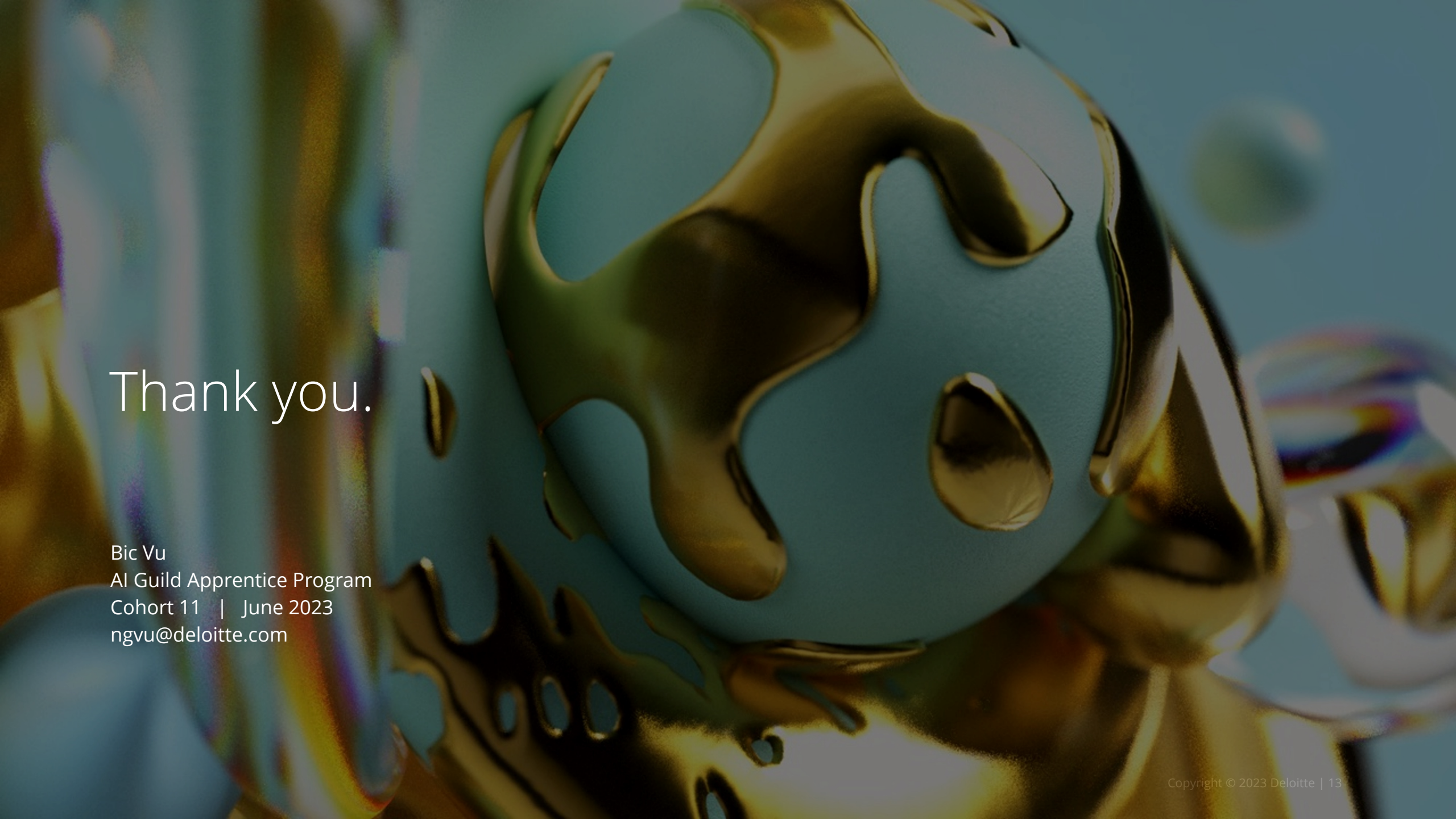
- Exposure to end to end regression model development, validation and testhing method
- Application of new libraries such as XGBoost, Ray Tune and AutoGluon

Growing Mastery

- Find data for home mortgage lenders who moved across different geographic locations
- Investigate how segregation and redlining overlaps with current study

Giving Back

- Contribute to initiatives or projects by bonding qualitative and quantitative research
- Enrich communities of practice by making ML practices accessible



Thank you.

Bic Vu
AI Guild Apprentice Program
Cohort 11 | June 2023
ngvu@deloitte.com

Data Notes

About the HMDA data

Source: CFPB
Year: 2021
Population: US
Geographic granularity: Census Tract

HMDA data dictionary

Metric	Defintion	Note
approval	Percent of approved vs denied applications in a census tract	Derived from "action_taken"
census_tract	US Census tract	
derived_race	Single aggregated race categorization derived from applicant/borrower and co-applicant/co-borrower race fields	
action_taken	The action taken on the covered loan or application	Filtered for: 1-loan originated 3 - application denied
loan_amount	The amount of the covered loan, or the amount applied for	
property_value	The value of the property securing the covered loan or, in the case of an application, proposed to secure the covered loan, relied on in making the credit decision	
income	The gross annual income, in thousands of dollars, relied on in making the credit decision, or if a credit decision was not made, the gross annual income relied on in processing the application	
debt_to_income_ratio	The ratio, as a percentage, of the applicant's or borrower's total monthly debt to the total monthly income relied on in making the credit decision	Binned into categories
tract_minority_ population_percent	Percent of census tract that is a minority population	
tract_to_msa_income_ percentage	Percentage of tract median family income compared to MSA/MD median family income	Binned into low, middle, high

About the Social Capital data

Source: Social Capital Study
Year: 2022
Population: US, Facebook users
Geographic granularity: Zip code

Social Capital Study Data

Metric	Defintion
ec_zip	**Economic connectedness (EC) – The degree to which low-income and high-income people are friends with each other.** Economic Connectedness: two times the share of high-SES (high income) friends among low-SES (low income) individuals, averaged over all low-SES individuals in the ZIP code.
nbhd_ec_zip	Economic connectedness calculated using only within-neighborhood friends.
ec_grp_mem_zip	Two times the share of high-SES friends among low-SES individuals averaged over all low-SES individuals in the ZIP code.
exposure_grp_mem_zip	The share of high-income people in the Facebook groups in which people participate. Mean exposure to high-SES individuals by ZIP code for low-SES individuals: two times the average share of high-SES individuals in individuals' groups, averaged over low-SES users.
clustering_zip	The average fraction of an individual's friend pairs who are also friends with each other. A metric of Cohesiveness: The degree to which social networks are fragmented into cliques
volunteering_rate_zip	The percentage of Facebook users who are members of a group which is predicted to be about 'volunteering' or 'activism' based on group title and other group characteristics. A measure of of Civic Engagement: Rates of volunteering and participation in community organizations
civic_organization_zip	The number of Facebook Pages predicted to be "Public Good" pages based on page title, category, and other page characteristics, per 1,000 users in the ZIP code. A measure of Civic Engagement.
num_below_p50	Number of children with below-national-median parental household income.

MODEL 1

XGBoost Regressor

Code snippet

```
model = XGBRegressor(
    n_estimators=300,
    max_depth = 5,
    random_state = seed,
    early_stopping_rounds=10
)
model = model.fit(X_train, y_train, eval_set=[(X_val, y_val)])
```

```
[0]  validation_0-rmse:0.23590
[1]  validation_0-rmse:0.16956
[2]  validation_0-rmse:0.12459
[3]  validation_0-rmse:0.09486
[4]  validation_0-rmse:0.07578
[5]  validation_0-rmse:0.06397
[6]  validation_0-rmse:0.05704
[7]  validation_0-rmse:0.05332
[8]  validation_0-rmse:0.05112
[9]  validation_0-rmse:0.05001
[10] validation_0-rmse:0.04938
[11] validation_0-rmse:0.04897
[12] validation_0-rmse:0.04876
[13] validation_0-rmse:0.04854
[14] validation_0-rmse:0.04836
[15] validation_0-rmse:0.04825
[16] validation_0-rmse:0.04829
[17] validation_0-rmse:0.04814
[18] validation_0-rmse:0.04811
[19] validation_0-rmse:0.04805
[20] validation_0-rmse:0.04800
[21] validation_0-rmse:0.04795
[22] validation_0-rmse:0.04798
[23] validation_0-rmse:0.04795
[24] validation_0-rmse:0.04789
[25] validation_0-rmse:0.04778
[26] validation_0-rmse:0.04775
[27] validation_0-rmse:0.04768
[28] validation_0-rmse:0.04765
[29] validation_0-rmse:0.04765
[30] validation_0-rmse:0.04761
[31] validation_0-rmse:0.04761
[32] validation_0-rmse:0.04758
[33] validation_0-rmse:0.04760
[34] validation_0-rmse:0.04754
[35] validation_0-rmse:0.04754
[36] validation_0-rmse:0.04754
[37] validation_0-rmse:0.04754
[38] validation_0-rmse:0.04755
[39] validation_0-rmse:0.04759
[40] validation_0-rmse:0.04760
[41] validation_0-rmse:0.04756
[42] validation_0-rmse:0.04751
[43] validation_0-rmse:0.04751
[44] validation_0-rmse:0.04749
[45] validation_0-rmse:0.04745
[46] validation_0-rmse:0.04738
[47] validation_0-rmse:0.04738
[48] validation_0-rmse:0.04736
[49] validation_0-rmse:0.04738
[50] validation_0-rmse:0.04738
[51] validation_0-rmse:0.04732
[52] validation_0-rmse:0.04721
[53] validation_0-rmse:0.04725
[54] validation_0-rmse:0.04724
[55] validation_0-rmse:0.04722
[56] validation_0-rmse:0.04721
[57] validation_0-rmse:0.04720
[58] validation_0-rmse:0.04724
[59] validation_0-rmse:0.04723
[60] validation_0-rmse:0.04728
[61] validation_0-rmse:0.04735
[62] validation_0-rmse:0.04736
[63] validation_0-rmse:0.04735
[64] validation_0-rmse:0.04734
[65] validation_0-rmse:0.04731
[66] validation_0-rmse:0.04732
[67] validation_0-rmse:0.04734
```


MODEL 2

XGBoost Regressor with Ray Tune

Code snippet

```
model = XGBRegressor(
    n_estimators=300,
    max_depth = 6,
    random_state = seed,
    early_stopping_rounds=10
)
model = model.fit(X_train, y_train, eval_set=[(X_val, y_val)])
```

```
Best model validation RMSE: 0.04585943734212902

Best model parameters: {
             'objective': 'reg:squarederror’,
             'eval_metric': 'rmse’,
             'n_estimators': 300,
             'max_depth': 6,
             'colsample_bytree': 0.8171484335790162,
             'lambda': 0.6673997656918179,
             'alpha': 0.03305586345304068}
```

[0]	validation_0-rmse:0.23443
[1]	validation_0-rmse:0.16778
[2]	validation_0-rmse:0.12254
[3]	validation_0-rmse:0.09210
[4]	validation_0-rmse:0.07235
[5]	validation_0-rmse:0.06027
[6]	validation_0-rmse:0.05334
[7]	validation_0-rmse:0.04941
[8]	validation_0-rmse:0.04708
[9]	validation_0-rmse:0.04585
[10]	validation_0-rmse:0.04525
[11]	validation_0-rmse:0.04492
[12]	validation_0-rmse:0.04468
[13]	validation_0-rmse:0.04448
[14]	validation_0-rmse:0.04436
[15]	validation_0-rmse:0.04431
[16]	**validation_0-rmse:0.04424**
[17]	validation_0-rmse:0.04431
[18]	validation_0-rmse:0.04431
[19]	validation_0-rmse:0.04431
[20]	validation_0-rmse:0.04435

[20]	validation_0-rmse:0.04435
[21]	validation_0-rmse:0.04439
[22]	validation_0-rmse:0.04425
[23]	validation_0-rmse:0.04424
[24]	validation_0-rmse:0.04429
[25]	validation_0-rmse:0.04430
[26]	validation_0-rmse:0.04429
[27]	validation_0-rmse:0.04426
[28]	validation_0-rmse:0.04427
[29]	validation_0-rmse:0.04427
[30]	validation_0-rmse:0.04419
[31]	validation_0-rmse:0.04417
[32]	validation_0-rmse:0.04420
[33]	validation_0-rmse:0.04419
[34]	validation_0-rmse:0.04424
[35]	validation_0-rmse:0.04423
[36]	validation_0-rmse:0.04427
[37]	validation_0-rmse:0.04425
[38]	validation_0-rmse:0.04419
[39]	validation_0-rmse:0.04424
[40]	validation_0-rmse:0.04424

MODEL 3

AutoGluon

Code snippet

```
from autogluon.tabular import TabularDataset, TabularPredictor

target = "approval"
metric = "rmse"
excluded_models = ['NN_TORCH', 'CAT', 'RF', 'FASTAI']

predictor = TabularPredictor(label=target).fit(
    train_data=df_train,
    excluded_model_types=excluded_models,
    time_limit=120,
    verbosity= 2,
    presets='best_quality’)

predictor.leaderboard(df_test, silent=True)

predictor.evaluate(df_test)
```

```
Best model: 'WeightedEnsemble_L3'


{'root_mean_squared_error': -0.03985582016238834,
'mean_squared_error': -0.001588486400816641,
'mean_absolute_error': -0.02876924000518339,
'r2': 0.7070154319401782,
'pearsonr': 0.8409075687940584,
'median_absolute_error': -0.021410344963759487}
```

