Key Recommendations

1. Offer flexible loan terms and origination charges.
2. Advertise flexible payment options for qualifying mortgage recipients to help ease the burden of paying off a home.
3. Design counseling services to guide homeowners in making informed decisions when purchasing a home.

Introduction

To gain deeper insights into consumer decisions when selecting a home mortgage lender, the data was organized into two sections: numerical and categorical predictor variables. This approach enables more effective processing and analysis of the data based on the nature of each predictor. Additionally, the variables can be grouped as needed, offering a clear understanding of the factors influencing consumer choices.

Numerical Predictors

Next, we will upload all datasets provided by Addition Financial for the competition[1]. The first dataset provides various types of loans and terms for 2023[2]. We will also upload the dataset for customer complaints[3] and the data for failed banks since 2000 and remove any missing values[4]. In the last two datasets, we will remove rows with missing values.

NPCA

The first process involves implementing normalized principal component analysis to look at the relationships between the predictor variables. When using PCA analysis we first find the covariance matrix for the variables. From there, we calculate the eigenvalues and eigenvectors to determine how much of the variance is explained by a component. Lastly, let Z be the principal components, X be the standardized variables, and P be the eigenvectors[5].

$$Z = XP$$

The first thing we need to do is normalize the variables in the HMDA loan dataset and calculate the principal components within R. Let's examine the

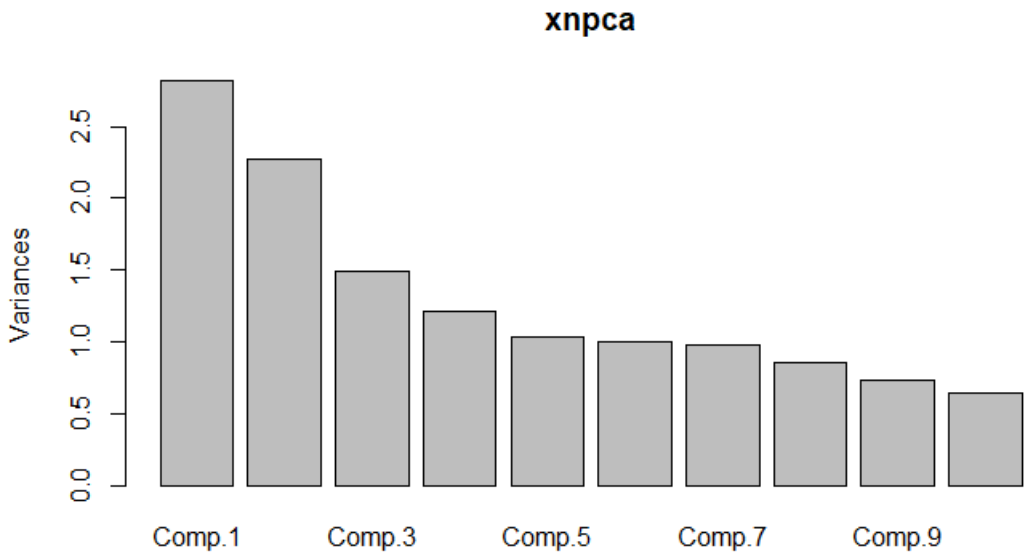

Figure 1: Variance explained with each principal component.

first two principal components since a majority of the variance is explained by these first two components.

```
Loadings:
                               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10
derived_msa_md                  0.108  0.155  0.479  0.356  0.216                0.408         0.379
loan_amount                    -0.197 -0.559  0.135                              0.187  0.170
interest_rate                                 0.139 -0.298  0.145         0.862 -0.178  0.276
origination_charges                   -0.281         0.419  0.146         0.381 -0.120 -0.738
loan_term                             -0.199 -0.146  0.589                      -0.510  0.532
property_value                 -0.188 -0.517  0.209 -0.147                       0.328  0.138
income                                                              0.981 -0.134 -0.110
tract_population               -0.496  0.156 -0.253                              0.165
tract_minority_population_percent      -0.101 -0.596        -0.110  0.107  0.193  0.488  0.168 -0.281
ffiec_msa_md_median_family_income      -0.294 -0.292 -0.442  0.286        -0.130 -0.229         0.489
tract_to_msa_income_percentage -0.277 -0.165  0.389 -0.114 -0.365               -0.203        -0.471
tract_owner_occupied_units     -0.534  0.207                0.155
tract_one_to_four_family_homes -0.477  0.274                0.254
tract_median_age_of_housing_units  0.232                    0.755        -0.175               -0.559
                               Comp.11 Comp.12 Comp.13 Comp.14
derived_msa_md                  0.490
loan_amount                    -0.190           0.717
interest_rate
origination_charges
loan_term                                      -0.167
property_value                 -0.227          -0.667
income
tract_population                0.147   0.742          -0.225
tract_minority_population_percent  0.334  -0.298           0.110
ffiec_msa_md_median_family_income  0.446  -0.119
tract_to_msa_income_percentage     0.542  -0.111          -0.152
tract_owner_occupied_units             -0.148           0.784
tract_one_to_four_family_homes     -0.181  -0.543          -0.535
tract_median_age_of_housing_units           0.105

                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13
SS loadings      1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000   1.000   1.000   1.000   1.000
Proportion Var   0.071  0.071  0.071  0.071  0.071  0.071  0.071  0.071  0.071   0.071   0.071   0.071   0.071
Cumulative Var   0.071  0.143  0.214  0.286  0.357  0.429  0.500  0.571  0.643   0.714   0.786   0.857   0.929
                Comp.14
SS loadings       1.000
Proportion Var    0.071
Cumulative Var    1.000
```

The tract population, tract owner-occupied units, and the number of one-to-four family homes show strong negative associations within the first principal component (PC1). The tract median age of housing units and tract MSA median family income have smaller values on PC1, contributing less to this component. We observe a strong negative association between property value and loan amount within the second principal component (PC2). Overall, while these variables show negative associations, the tract population, tract owner-occupied units, one-to-four family homes, property value, and loan amount are the most influential within the data.

Factor Analysis

We will also look at the factors for the numerical variables for the loans. When we use factor analysis, we want to reduce the number of variables using factors[5].

$$x_j = \sum_{l=1}^{k} q_{jl} f_l + \mu_j, j = 1, \ldots, p.$$

We will plot the first two factors with the specific variable names within the scaled numerical variables. Looking at the factors plotted, it appears that the

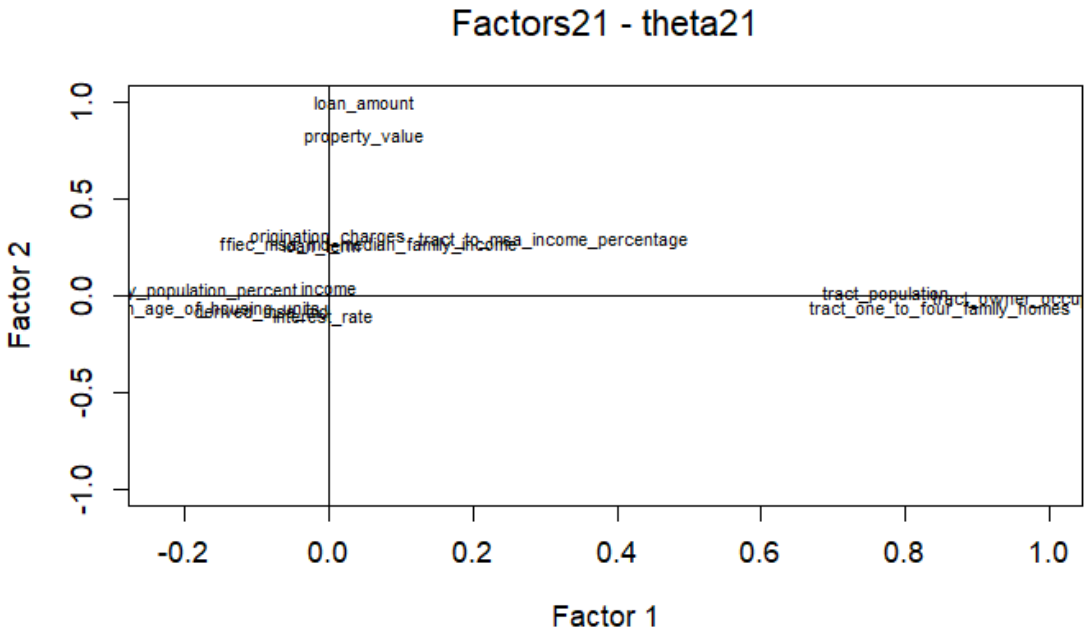

Figure 2: Factors for HMDA data.

tract population, tract one-to-four family homes, and tract owner occupancy have a positive association in the first factor while the loan amount and property value have the most significant association in the second factor. When comparing these results to those from NPCA, we find that the factors align with the principal components, with the same variables being grouped together in each factor.

Newton-Raphson

The next piece of information we will look at from the loans dataset is what variables allow people to borrow larger amounts. This may give insight into which demographics will be most beneficial to attract when advertising. Thus, let the loan amount be our target variable and all other numerical variables in the scaled dataset will be our predictors with the first 1000 observations serving as the training data/ Lets now use the Newton-Raphson method to find the maximum likelihood estimator. This involves continually updating β_{t+1} using the following equation.

$$\beta_{t+1} = \beta_t - H^{-1}\Delta f(\beta_t)$$

The Hessian matrix (H) is the second derivative of the maximum likelihood function for a normal distribution.

$$\frac{\partial}{\partial \beta}(\Delta f(\beta) = X^T X$$

The gradient $\Delta f(\beta_t)$ is the first derivative of the maximum likelihood function for a normal distribution.

$$\frac{\partial}{\partial \beta} f(\beta) = X^T(Y - X\beta)$$

Thus, in the end, we have the formula:

$$\beta_{t+1} = \beta_t - (X^T X)^{-1}(-X^T(Y - X\beta_t))$$

Next, we will define the initial values for the coefficient β_{t+1} and β_t as 1 and 0 respectively. To begin the algorithm, we will define a loop in R that will continually adjust the value of β_{t+1} until it reaches convergence or $\beta_{t+1} - \beta_t < 0.000001$. Within this loop, the function will update β_t with β_{t+1} as well as calculate the Hessian and gradient matrices.

	0.0359
derived_msa_md	0.0409
interest_rate	0.0440
origination_charges	0.1539
loan_term	0.1473
property_value	0.6551
income	45.2324
tract_population	-0.0337
tract_minority_population_percent	-0.0280
ffiec_msa_md_median_family_income	0.0250
tract_to_msa_income_percentage	0.0084
tract_owner_occupied_units	0.0439
tract_one_to_four_family_homes	-0.0149
tract_median_age_of_housing_units	-0.0019

The results indicate that income has the most significant impact on the loan amount, which is expected, as a person's income directly influences how much they can borrow. The next most influential predictors are property value, origination charges, and loan term. These findings suggest that Addition should target borrowers with higher property values and place more emphasis on advertising loan terms and origination charges.

Variable Shrinkage

Let's now compare the results with Lasso, Ridge, and Elastic-Net variable shrinkage. The first method we will use is Lasso and the variables will be standardized.

Lasso

When we use Lasso[5], we look to estimate the coefficient $\hat{\beta}$ with a tuning parameter $\lambda \geq 0$ such that

$$argmin_{\beta}[\Sigma_{i=1}^{n}(y - x^{T}\beta)^{2} - \lambda\Sigma_{i=1}^{n}|\beta_{j}|].$$

Let's now find the optimal λ to use within our analysis. We will use k-fold cross validation to find which value of λ has the lowest prediction error[7].

```
[1] 0.0074
```

Implementing the Lasso method, the β values will also be sorted from least to greatest.

```
tract_minority_population_percent -0.0315
...2 0.0000
tract_population 0.0000
tract_one_to_four_family_homes 0.0000
tract_median_age_of_housing_units 0.0000
tract_owner_occupied_units 0.0022
tract_to_msa_income_percentage 0.0064
ffiec_msa_md_median_family_income 0.0153
derived_msa_md 0.0235
interest_rate 0.0258
(Intercept) 0.0433
loan_term 0.1419
origination_charges 0.1490
property_value 0.6514
income 44.5521
```

Income serves as the most significant predictor of loan amount as well as property value, origination charges, and loan term. All other variables have minimal or no effect on the loan amount.

Ridge

With the Ridge technique, we want to minimize β such that

$$min_B \sum_{j=1}^{n} (y_i - X\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2.$$

Again, we begin by finding the optimal value of lambda using k-fold cross validation.

```
[1] 0.013
```

Next, we will use the glmnet function to calculate the values of β.

```
tract_population -0.0315
tract_minority_population_percent -0.0273
tract_one_to_four_family_homes -0.0115
tract_median_age_of_housing_units -0.0029
...5 0.0000
tract_to_msa_income_percentage 0.0124
ffiec_msa_md_median_family_income 0.0263
(Intercept) 0.0383
derived_msa_md 0.0387
tract_owner_occupied_units 0.0393
interest_rate 0.0432
loan_term 0.1452
origination_charges 0.1529
property_value 0.6427
income 45.9669
```

Same as Lasso, the income remains the most influential predictor followed by the property value, origination charges, and loan term.

Elastic-Net

In the final method, we will use Elastic Net feature select. In Elastic Net, we again use standardized variables and look to estimate $\hat{\beta}$ with $\lambda \geq 0$ such that

$$argmin_{\beta}[(2n)^{-1}\Sigma_{i=1}^{n}(y - x^{T}\beta)^{2} + \lambda P_{\alpha}(\beta)].$$

Here, P_{α} is the elastic-net penalty. We will again use k-fold cross validation to find the value of λ with the lowest prediction error.

```
[1] 0.013
```

We will now implement the elastic-net function within R to find the values of the coefficients and sort them.

```
  tract_minority_population_percent -0.0313
...2 0.0000
tract_population 0.0000
```

```
tract_one_to_four_family_homes 0.0000
tract_median_age_of_housing_units 0.0000
tract_owner_occupied_units 0.0030
tract_to_msa_income_percentage 0.0087
ffiec_msa_md_median_family_income 0.0167
derived_msa_md 0.0243
interest_rate 0.0273
(Intercept) 0.0439
loan_term 0.1413
origination_charges 0.1491
property_value 0.6455
income 45.0145
```

Once again, the income serves as the most significant variable followed by the property value, origination charges, and loan term. The tract population, tract one-to-four family homes, tract median age of housing units, tract owner occupied units, tract to msa income percentage 0.0087 was reduced to 0 or close to 0 suggesting it as little to no influence on the loan amount.

Categorical Predictors

Feature Selection for HMDA

We will now look at the non-numeric variables within the HMDA dataset[2]. Beginning with isolating the remaining variables not compared in the previous section, we put these into a new dataframe along with the loan amount as the target variable. Next, all entries in each column that make up less than 5% of the column gets converted into other in order to reduce the number of dummy variables created. Once the dummy variables are created, a generalized linear model is fit with all of the categorical predictors. Finally this model gets uploaded into stepwise forward selection until the best combination of variables is found. In stepwise forward selection, a variable is added until it is not statistically significant to add any more variables [6]. After running this algorithm, the top three variables that had the greatest positive influence on the borrowing amount all included the occupancy type for the loan. Thus, whether the mortgage is for a house that is primary residence or a house that will be used as an investment property, it may lead to higher loan amounts.

Mortgage Complaints

Next, we will look at the complaints data set to determine which complaints occur most frequently [3]. For this, the data will be restricted to include only the individuals with a mortgage loan and all rows with an "N/A" value will be omitted. Creating a bar graph to showcase the most common issues in the dataset, we will look at which complaints influence the most customers. The

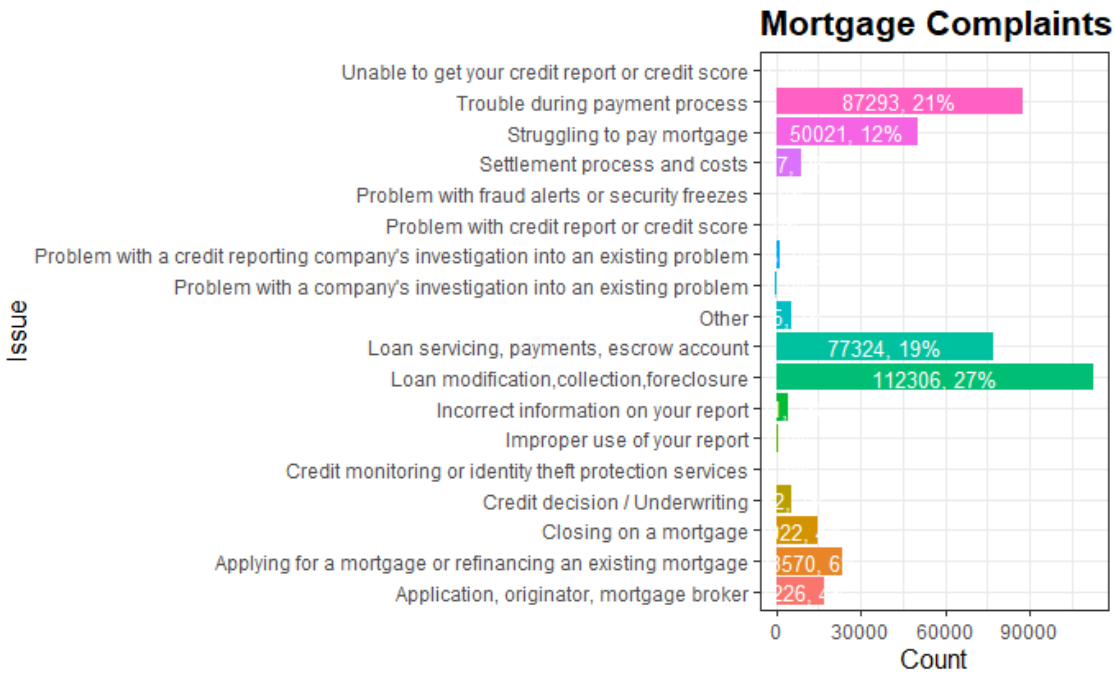

Figure 3: Count of mortgage complaints.

most prevalent issues appear to be related to "loan modifications, collections, and foreclosures," followed by "payment process difficulties" and "loan servicing, payments, and escrow accounts." This suggests that a large number of individuals faced challenges with making timely payments, which subsequently led to collections or foreclosure. Thus, it would prove beneficial for a financial institution to offer more flexible payment options for customers and offer services to counsel home owners in the best options for buying a house.

Conclusion

Using variable shrinkage methods as well as carefully selected data exploration and visualization, the key factors influencing consumer's decisions gets revealed. From the non-numeric data we can see that allowing for flexible payment options and counseling on affording mortgages would lead to less consumer complaints. This leads to ore customers choosing your institution when applying for a mortgage. Furthermore, from the numeric predictor variables, offering flexible loan terms and origination charges have a significant influence on the loan amount requested. Offering these two will lead to higher loan amount which leads to more interest and the customer returning for other kinds of loans.

References

[1] Addition Financial. (2024, October 20). Addition financial competition 2024-2025. Statistics and Data Science. https://sciences.ucf.edu/statistics/addition-financial-competition-2024-2025/

[2] FFIEC. (n.d.). HMDA - Home Mortgage Disclosure act. https://ffiec.cfpb.gov/data-publication/snapshot-national-loan-level-dataset/2023

[3] Consumer complaint database. Consumer Financial Protection Bureau. (n.d.). https://www.consumerfinance.gov/data-research/consumer-complaints/#download-the-data

[4] Publisher Division of Insurance and Research. (2020, November 12). Federal Deposit Insurance Corporation - FDIC Failed Bank list. Catalog. https://catalog.data.gov/dataset/fdic-failed-bank-list

[5] Hardle, W., & Simar, L. (2015). Applied Multivariate Statistical Analysis. Springer Berlin Heidelberg.

[6] Forward selection. Forward Selection - an overview | ScienceDirect Topics. (n.d.). https://www.sciencedirect.com/topics/computer-science/forward-selection#:~:text=2.2%20Stepwise%20selection%20methods&text=Forward%20selection%20(FS)%3A%20Starting,sum%20of%20squares%20(RSS).