Access to Credit and Financial Well-Being

March 8, 2024 Shaun Campbell COSC 6520

1 Introduction

The ability to access credit, such as borrowing money or getting a credit card, is a factor that may influence an individual's financial well-being. According to a study done by the CFPB, having a credit card, one's credit card limit, and number of credit accounts are all positively correlated with one's financial well-being[1].

There also exists a large disparity in credit access among different demographics. For example, Black Americans are 2.2 times less likely to have access to prime rates versus the overall American population and Hispanic Americans are 1.8 times less likely[2], and those aged 18-25 have an average credit score almost 10% lower than those 58-76[3].

Understanding how demographic characteristics associate with the ability to get credit is an important step to addressing these disparities. Uncovering the demographic factors that lessen credit access would provide a starting point for investigating potential systemic factors such as biased lending practices. It would also provide reason to address other potential causes for disparities such as credit score differences among different demographics.

Since certain demographics are clearly underserved by lenders, and if credit access can in fact improve financial well-being, then these demographics are being harmed by current lending practices. Establishing that the ability to access credit directly influences financial well-being would provide reason for developing programs to improve credit access to underserved populations. Programs such as rent reporting to help Americans build credit[2] and government-subsidized lending bodies are potential solutions that could expand credit access to those who are underserved.

2 Statistical Methodology

The first objective is to develop a model that can determine an individual's likelihood of obtaining credit based on their demographic information. The dataset used for this task is the 2018 Loan Application Register ("LAR"), and a summary of the dataset is included in Appendix 1.

The second dataset is the Financial Well Being Survey ("FWBS") data provided by the CFPB. This survey was performed in 2017 and the CFPB determined a financial well-being score based on responses. The dataset also contains demographic information for the respondents and is outlined in Appendix 2.

The LAR data is first reduced to mortgage applications associated with a primary-residence, individual home purchase, and approval/denial is used as a binary target variable. The FWBS data contains only categorical data, so any numerical variables in the LAR dataset are binned appropriately to match the levels of the FWBS variables. Some feature engineering is performed: the property value minus the loan amount is used to approximate the amount of the down payment; VA mortgages are used to indicate military status; having a co-applicant is used as a proxy for marital status.

The data is then split into train, validation, and test sets. The data is imbalanced, containing about 87.5% approved applications, so the train set is resampled to contain a balanced number of each target variable. Under sampling is used to keep the size manageable and reduce computational expense. After under sampling, there are around 1.7 million records across the three sets distributed 50/25/25 for train, validation, and test. The validation set and test set maintain the imbalanced class levels.

A naïve Bayes model is trained on the train set using 5-fold cross validation. Naïve Bayes is quick and efficient and is used here to get an idea of the viability of the classification problem. With an F1 score of 0.769 on the validation set and an AUC of 0.687, the approach lends evidence that classification based on demographics is possible. Output of the naïve Bayes model fitting is included in Appendix 3.

A classification tree model is evaluated next because this method is intuitive for making credit approval decisions and has been known for its use in the banking industry. The full tree is first generated and the best-pruned tree is selected as the tree with the least depth while maintaining a cross validation error within 1 standard deviation of the minimum error tree. The best pruned tree has a depth of 83 and 337 splits. The performance of this tree and trees with depths of 10 and 50 are evaluated on the validation set. The best-pruned tree performs best with an F1 score of 0.802 and an AUC of 0.710. Output of the classification tree fitting is included in Appendix 4.

Ensemble methods are then explored to see if performance can be further improved with tree-based classifiers and to obtain feature importance metrics. With k being the number of variables randomly selected for the weak learners, the number of trees in the ensemble is held constant at 100 trees and different values of k are tested. k = 3 performs best with an F1 score is 0.805 and AUC of 0.711 on the validation set. Using the underlying distribution of 12.5% rejection rate, the k = 3 model has an F1 score of 0.869. Output of the ensemble tree fitting is included in Appendix 5.

The best classifier for each technique is evaluated on the test set to select the final model, and performance metrics are displayed in Table 1. The random forest model with k = 3 performs best with an F1 score of 0.804 and an AUC of 0.713. The basic classification tree performs almost just as well as the random forest model, and would have the benefit of improved interpretability and simplicity. However, with a high depth value (83) and number of splits (337), overfitting and the model's ability to generalize to unseen data are of concern. Due to random forest models being less prone to overfitting, the k =3 random forest with a 0.125 cutoff is selected for the final model. Additional output can be found in Appendix 6.

Algorithm	Parameters	F1	Sensitivity	Specificity	Precision	AUC
Naïve Bayes		0.767	0.657	0.614	0.923	0.686
Classification Tree	depth = 83	0.801	0.703	0.625	0.929	0.711
Random Forest	k = 3	0.804	0.708	0.623	0.929	0.713
Random Forest, 0.125 cutoff	k = 3	0.869	0.827	0.473	0.917	

Table 1: Model Performance Metrics on the Test Set

3 Results

The F1 score of the final model (0.869) means the model has a good balance between precision and sensitivity. Since the data is highly imbalanced towards approvals, the high precision (0.917) is not necessarily indicative of good performance. The decile-wise lift chart does, however, show that for the top 60% of positive probabilities the model provides lift above random guessing (Figure 1). The ROC curve (Figure 2) and corresponding AUC demonstrate that the model has fair performance in terms of discriminating between approvals and denials. It can thus be concluded that a prospective borrower's demographics, such as where they live, their race/ethnicity, and income are associated with their chances of being approved for a mortgage, however demographic information alone is far from deterministic.

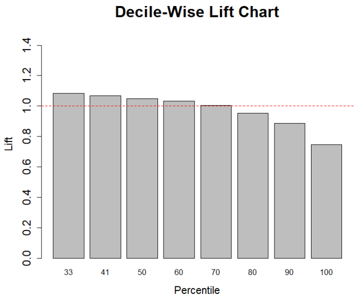


Figure 1: Decile-Wise Lift Chart of Random Forest Model

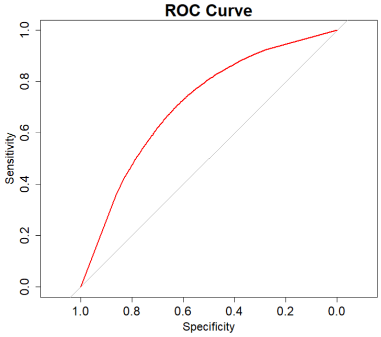


Figure 2: ROC Curve of Random Forest Model

The feature importance of the random forest model shows which demographic factors are most important to one's probability of obtaining a home loan (Figure 3). Income is the most important discriminative factor, followed by the down payment made. This makes sense because someone with a higher income would assumedly have a better ability to make loan payments, and a higher down payment means less money needs to be recovered from the borrower. Race and/or ethnicity is the next most important factor. This could mean that certain racial and ethnic groups are trailing others on average in terms of non-demographic factors that determine loan approval such as credit score. It could also mean that lending algorithms have implicit bias and is a reason for further investigation. Military status is fourth-most important, which could be attributed to veterans' and military members' ability to obtain VA loans. Age, the region of the county one lives in, whether they live in a metro area, marital status, and sex are all relatively unimportant compared to the other factors.

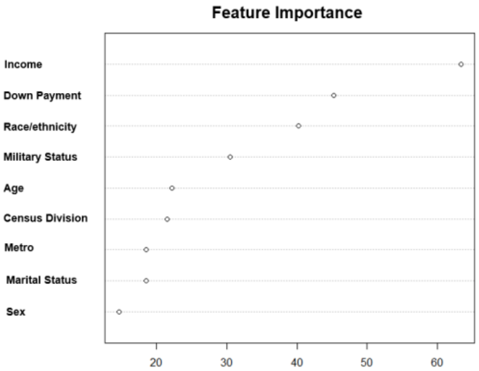


Figure 3: Feature Importance in Random Forest Model

With an understanding of which demographic factors are important in determining the probability one is approved for a mortgage, the probability of obtaining a home loan can be used as a proxy for ability to obtain credit in general. The assumption made here is that algorithms for credit decisions are all similar, and therefore demographic information would associate generally with credit approval.

Next, the FWBS data is transformed to have the same format as the LAR data. Amount of money in savings is used as a proxy for down payment because theoretically this is the maximum amount one could put towards a down payment. Some indicator variables are also used for further analysis: whether the individual has a credit card and whether they own a home. The random forest algorithm is then applied to the dataset to a create a variable for loan approval probability as a proxy of that individual's ability to access credit based on their demographic information.

After approval probability is added to the FWBS data set, the relationship between ability to access credit and financial well-being is investigated. There is a positive relationship between loan approval probability and financial well-being score (Figure 4). When regressing financial well-being score on approval probability (Appendix 7), approval probability has a positive coefficient and is individually significant. The R^2 is 0.19 meaning 19% of the variance in financial well-being is explained by credit approval probability. With credit approval probability acting as a reflection of demographic information, this translates to mean that 19% of variance in financial well-being is explained by demographics.

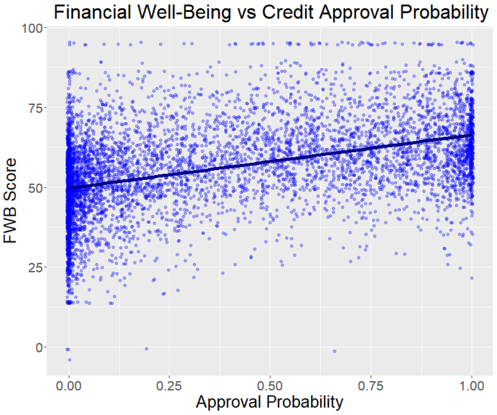


Figure 4: Relationship Between Approval Probability and FWB Score

However, using just the two most important features – income and down payment – raises the adjusted R^2 to 0.347. This means that income and savings amount explain almost double the variance in financial well-being than the aggregate credit approval probability based on demographic information.

Next, the relationship between access to credit and financial well-being is explored across different populations. First the data is broken down into two subsets: having credit (own a home or have a credit card) and do not have credit (do not own a home and do not have a credit card). It can be seen in Figure 5 that those with credit have, on average, a higher financial well-being than those who do not across all levels of approval probability, and difference in means is statistically significant (Appendix 8). This means that given two people with the exact same demographics, the one with credit will on average have a higher financial well-being score than the one without. It could be theorized that obtaining credit plays a role in causing an increase in financial well-being, but due to the high number of possible confounding variables this conclusion cannot be made with any confidence. A simple explanation to the opposite is that individuals who already have a high financial well-being have a higher credit score on average and thus have a better ability to obtain credit approval.

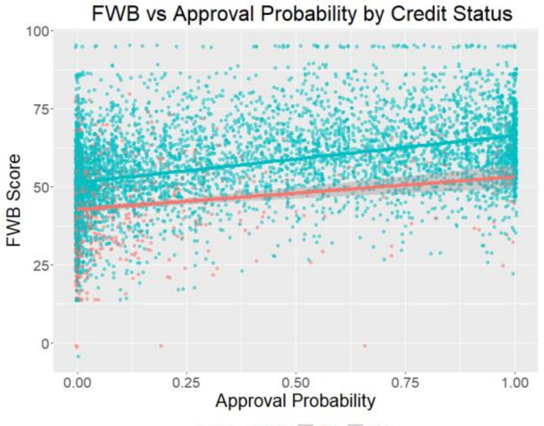


Figure 5: FWB Score and Approval Probability by Credit Status

The data is then broken down by race/ethnicity – the scatter plot is included in Appendix 9. The positive relationship between credit access and financial well-being score holds true across the four race/ethnicities explored. It appears that Black, Non-Hispanic individuals have the most to gain in terms of financial well-being when approval probability increases, as the slope of the regression line for this group is steeper than the regression lines for the other groups. With an increase of 10% approval probability, the financial well-being score of Black, Non-Hispanic people will increase on average by 2.27, versus 1.66, 1.44, and 1.60 for White-Non-Hispanic, Hispanic, and Other, Non-Hispanic respectively (Appendix 9).

Finally, the relationship among different income levels is investigated. The trend is notable for individuals in the lowest income level (<$40,000). These individuals have an average FWB score of 48.8 versus 64.1 for those making over $150,000 (Appendix 10), but a 10% increase in approval probability results in an average financial well-being score gain of 3.06 versus 1.30 for those making over $150,000. However, the R^2 value for this group is very low with only 3% of the variance in financial well-being score being explained by variance in approval probability (Appendix 11).

4 Alternative Approaches

Techniques such as k-nearest neighbors and principal component analysis were not used because these methods were not designed to handle categorical variables (although not explored, one-hot encoding could be used to evaluate a KNN model) and methods such as regression trees were not used because the target variable is a binary class. Including more variables in the original LAR classification models could have resulted in better performance, but the objective was to focus on the predictive ability of solely demographic information. Aggregating an individual's demographic information to derive a proxy for their ability to access credit adds a valuable dimension to the financial well-being data for analysis and interpretation rather than just focusing on the survey data alone. The methods in this study were an attempt to circumnavigate the lack of publicly available longitudinal financial well-being data and allow the relationship between credit access and financial well-being to be investigated without before and after comparisons on an individual level.

5 Conclusions

It is important to study specific demographic factors that contribute to disparities in credit access so the disparities can be addressed. This investigation found that after income and down payment amount, race/ethnicity is the third most import demographic factor that contributes to credit approval. For example, Black, non-Hispanic individuals have an approval probability 62% lower than White, Non-Hispanics and have an average financial well-being score 3% lower (Appendix 10). The reasons behind this fact should be further researched: whether the differences are due to biased lending algorithms and/or the causes behind other underlying disparities in different racial and ethnic groups – such as Black, Non-Hispanics having an average credit score 8% lower than White, Non-Hispanics[4].

Establishing that the ability to obtain credit directly and positively affects financial well-being would provide reason to establish programs and new regulations that help Americans who are underserved by lenders access credit. This study finds evidence of a positive relationship between credit approval probability and financial well-being; however, the directionality of the relationship cannot be determined. The findings could mean that a greater ability to access credit does in fact tend to improve one's financial well-being, but could also simply mean those with a high financial well-being tend to have similar demographic information, for example income, and thus a higher probability of being approved. With the attempt to circumvent lack of available longitudinal data on financial well-being and credit proving inconclusive, it is recommended that organizations such as the CFPB begin collecting this data so longitudinal studies can establish the true relationship between credit access and financial health.

6 References

1. Nagypál , É., & Tobacman , J. (2019, September 17). *New report explores the relationship between financial well-being and the contents of and engagement with credit reports*. Consumer Financial Protection Bureau. https://www.consumerfinance.gov/about-us/blog/new-report-explores-relationship-between-financial-well-being-contents-engagement-credit-reports/
2. Cruz-Martinez, G. (2022, February 9). *Millions could get a leg up financially if credit scores included alternative data*. Yahoo! Finance. https://finance.yahoo.com/news/credit-scores-alternative-data-160339352.html?guccounter=1
3. Crace, M. (2023, September 29). *Here's The Average Credit Score By Age In The US*. Rocket Money. https://www.rocketmoney.com/learn/debt-and-credit/what-is-the-average-credit-score-by-age
4. Karl, S. (2024, January 17). *Average credit scores by Race*. Investopedia. https://www.investopedia.com/average-credit-scores-by-race-5214521

Appendix 1

Summary of the Loan Application Register ("LAR") data. Demographic information was not added to the LAR until 2018 so this year is chosen as the closest possible to the 2017 FWBS.

Available here: https://ffiec.cfpb.gov/data-publication/three-year-national-loan-level-dataset/2018

Data dictionary: https://ffiec.cfpb.gov/documentation/publications/loan-level-datasets/lar-data-fields

Variable	Description	Type
action_taken	Action taken on loan application (approved, denied)	Categorical
loan_purpose	The purpose of loan (home purchase, refi, etc)	Categorical
open_end_line_of_credit	Whether application is for an open end line of credit	Categorical
business_or_commercial_purpose	Whether application is for primarily a business or commercial purpose	Categorical
occupancy_type	Occupancy type for dwelling (principal residence, second residence, etc)	Categorical
applicant_age	The age of the applicant (binned)	Categorical
county_code	State-county FIPS code	Categorical
state_code	Two-letter state code	Categorical
derived_ethnicity	Ethnicity derived from applicant and co-applicant	Categorical
derived_race	Race derived from applicant and co-applicant	Categorical
derived_sex	Sex derived from applicant and co-applicant	Categorical
derived_loan_product_type	Loan product type and lien status (conventional first lien, VA subordinate lien, etc)	Categorical
co_applicant_credit_score_type	Version of credit scoring model used for co-applicant	Categorical
income	Gross annual income	Numerical
loan_amount	Amount of loan	Numerical
property value	Value of the property	Numerical

Appendix 2

Summary of the Financial Well-Being Survey ("FWBS") data

Available here: https://www.consumerfinance.gov/data-research/financial-well-being-survey-data/

Data dictionary: https://files.consumerfinance.gov/f/documents/cfpb_nfwbs-puf-codebook.pdf

Variable	Description	Type
PPMSACAT	MSA Status (Non-metro, metro)	Categorical
PPREG9	Census Division	Categorical
PPETHM	Race/Ethnicity	Categorical
PPGENDER	Gender	Categorical
agecat	Age (binned)	Categorical
PPINCIMP	Household income (binned)	Categorical
HOUSERANGES	Amount paid for home each month (binned)	Categorical
SAVINGSRANGES	Amount in savings (binned)	Categorical
MANAGE1_3	Frequency credit card is paid off in full each month	Categorical
REJECTED_1	Applied for credit and was turned down	Categorical
VALUERANGES	What respondent thinks their home would be worth if sold	Categorical
MORTGAGE	How much is currently owed on home	Categorical
PPMARIT	Marital Status	Categorical
MILITARY	Current/former, spouse of, dependent of, member of US Armed Forces	Categorical
PCTLT200FPL	County percent less than 200% of poverty level	Categorical
ENDSMEET	Difficulty covering monthly bills	Categorical
FWBscore	Financial well-being scale score	Categorical

Appendix 3

Output of Naïve Bayes Model Fitting and Evaluation on Validation Set

Naïve Bayes	F1	Sensitivity	Specificity	Precision	Recall	AUC
cutoff = 0.5	0.769	0.659	0.615	0.923	0.659	0.687
cutoff = 0.125	0.931	0.983	0.095	0.884	0.983	

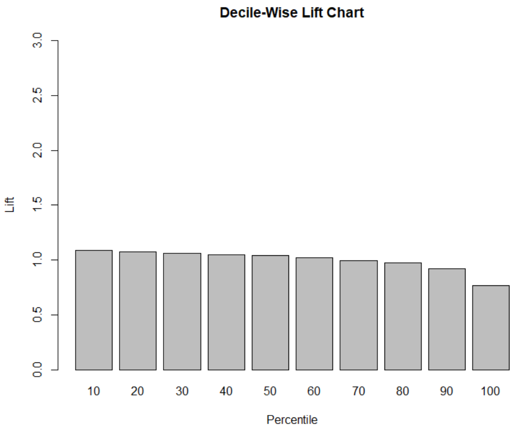
Performance of Naïve Bayes models on Validation Set

	Reference	
Prediction	0	1
0	33132	129134
1	20784	249083

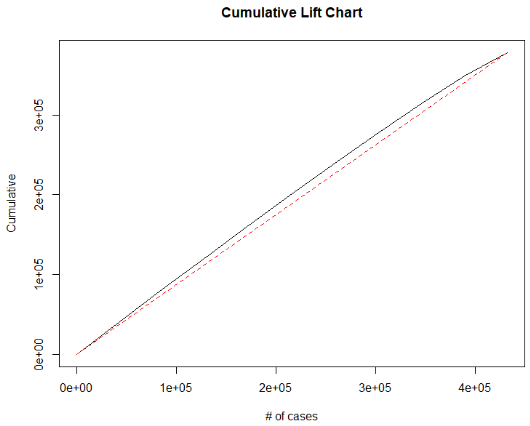
Confusion Matrix, Validation Set, cutoff = 0.5

	Reference	
Prediction	0	1
0	5137	6556
1	48779	371661

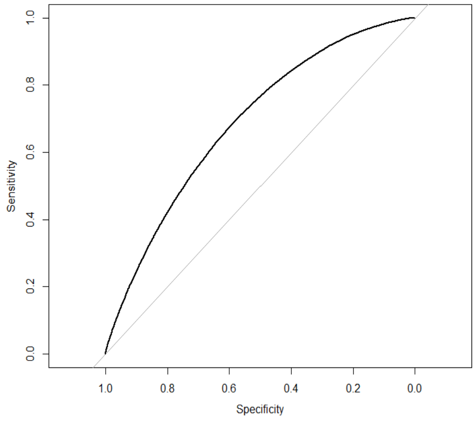
Confusion Matrix, Validation Set, cutoff = 0.125



Decile-Wise Lift chart on Validation Set



Cumulative Lift Chart on Validation Set



ROC Curve on Validation Set

Appendix 4

Output of Classification Tree Model Fitting and Evaluation on Validation Set

Classification Tree	F1	Sensitivity	Specificity	Precision	Recall	AUC
depth = 10	0.817	0.733	0.580	0.924	0.733	0.689
depth = 50	0.806	0.712	0.613	0.928	0.712	0.704
depth = 83	0.802	0.705	0.623	0.929	0.705	0.710

Performance of Classification Tree Models on Validation Set

	Reference	
Prediction	0	1
0	31257	101116
1	22659	277101

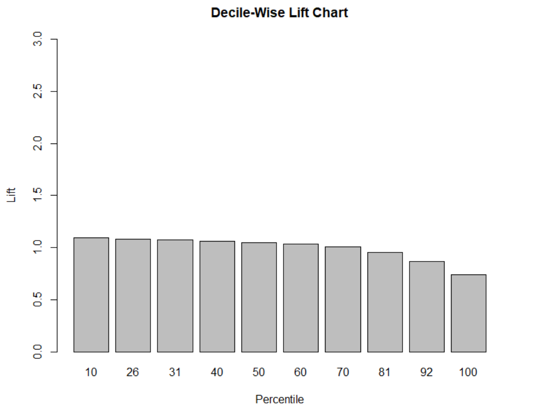
Confusion Matrix, Validation Set, depth = 10

	Reference	
Prediction	0	1
0	33024	108822
1	20892	269395

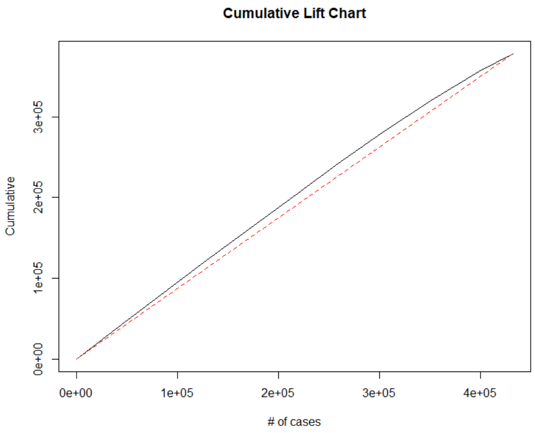
Confusion Matrix, Validation Set, depth = 50

	Reference	
Prediction	0	1
0	33596	111635
1	20320	266582

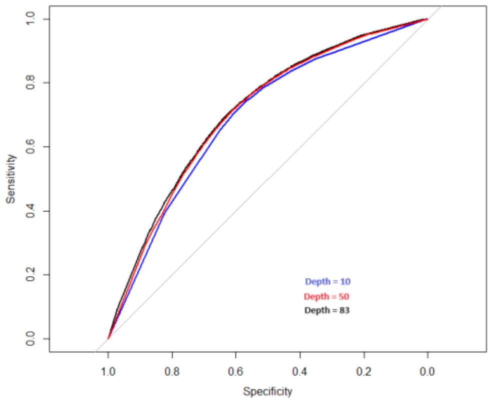
Confusion Matrix, Validation Set, depth = 83



Decile-wise lift chart on validation set (depth=83)



Cumulative lift chart on validation set (depth=83)



AUC curves on validation set for different depths

Appendix 5

Output of Ensemble Model Fitting and Evaluation on Validation Set

Random Forest	F1	Sensitivity	Specificity	Precision	Recall	AUC
k = 1	0.809	0.720	0.585	0.924	0.720	0.701
k = 2	0.804	0.719	0.606	0.927	0.719	0.709
k = 3	0.805	0.710	0.620	0.929	0.710	0.711
k = 5	0.799	0.702	0.618	0.928	0.702	0.706
k = 9 (bagging)	0.792	0.691	0.616	0.927	0.691	0.699

Performance of Ensemble Tree Models on Validation Set

	Reference	
Prediction	0	1
0	31538	106087
1	22378	272130

	Reference	
Prediction	0	1
0	32663	106453
1	21253	271764

Confusion Matrix, Validation Set, k=1

Confusion Matrix, Validation Set, k=2

	Reference	
Prediction	0	1
0	33421	109535
1	20495	268682

	Reference	
Prediction	0	1
0	33336	112816
1	20580	265401

Confusion Matrix, Validation Set, k=3

Confusion Matrix, Validation Set, k=5

	Reference	
Prediction	0	1
0	33208	116808
1	20708	261409

Confusion Matrix, Validation Set, k=9 (bagging)

Appendix 6

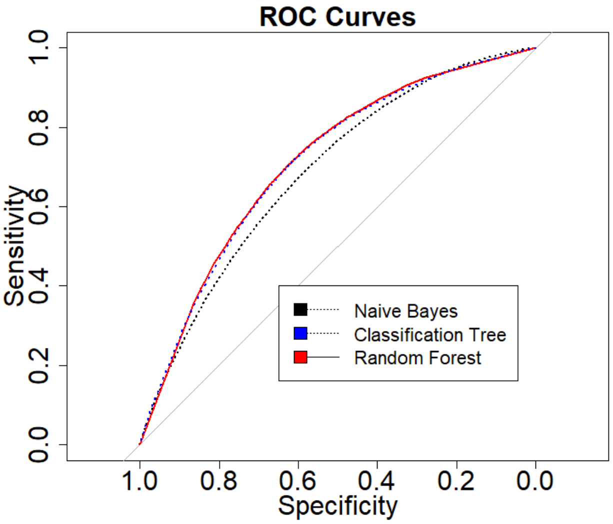
Evaluation of Models on Test Set

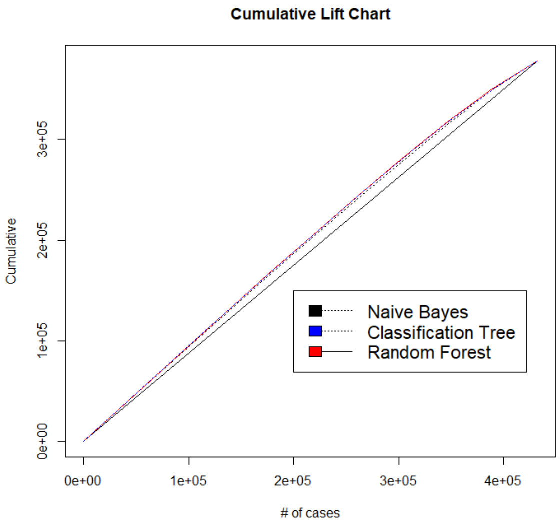
```
                Reference
Prediction           0          1
         0       33110     129827
         1       20805     248389
      Reference
Prediction           0          1
     0       33702     112301
     1       20213     265915
```

Confusion Matrix, Test Set, Naïve Bayes *Confusion Matrix, Test Set, Classification Tree*

```
                Reference
Prediction           0          1
         0       22587     110503
         1       20328     267713
      Reference
Prediction           0          1
     0       22505      65557
     1       28410     312659
```

Confusion Matrix, Test Set, Random Forest *Confusion Matrix, Test Set, Random Forest, cutoff = 0.125*





ROC Curves on Test Set *Cumulative Lift Chart on Test Set*

Appendix 7

Linear Regression Results

	Model 1	Model 2
Intercept	49.73 (<2e-16)	49.938 (<2e-16)
Prob_approved	16.70 (<2e-16)	
Income $150,000 or more		1.726 (0.001)
Income $75,000 to $99,999		-0.360 (0.475)
Income $60,000 to $74,999		-1.377 (0.015)
Income $50,000 to $59,999		-1.228 (0.047)
Income $40,000 to $49,999		-2.388 (0.000)
Income Less than $39,999		-5.9861 (<2e-16)
Down Payment $75,000 or more		19.287 (<2e-16)
Down Payment $20,000-74,999		14.551 (<2e-16)
Down Payment $5,000-19,999		10.379 (<2e-16)
Down Payment Unknown		9.923 (<2e-16)
R^2	0.1901	0.3783
Adj R^2	0.1899	0.3473
F-Stat	1500	341.1

Regression Results for Linear Regression Models. Model 1 Regresses FWB score on Probability of Approval, Model 2 Regresses FWB score on Income and Down Payment (Amount in Savings)

Appendix 8

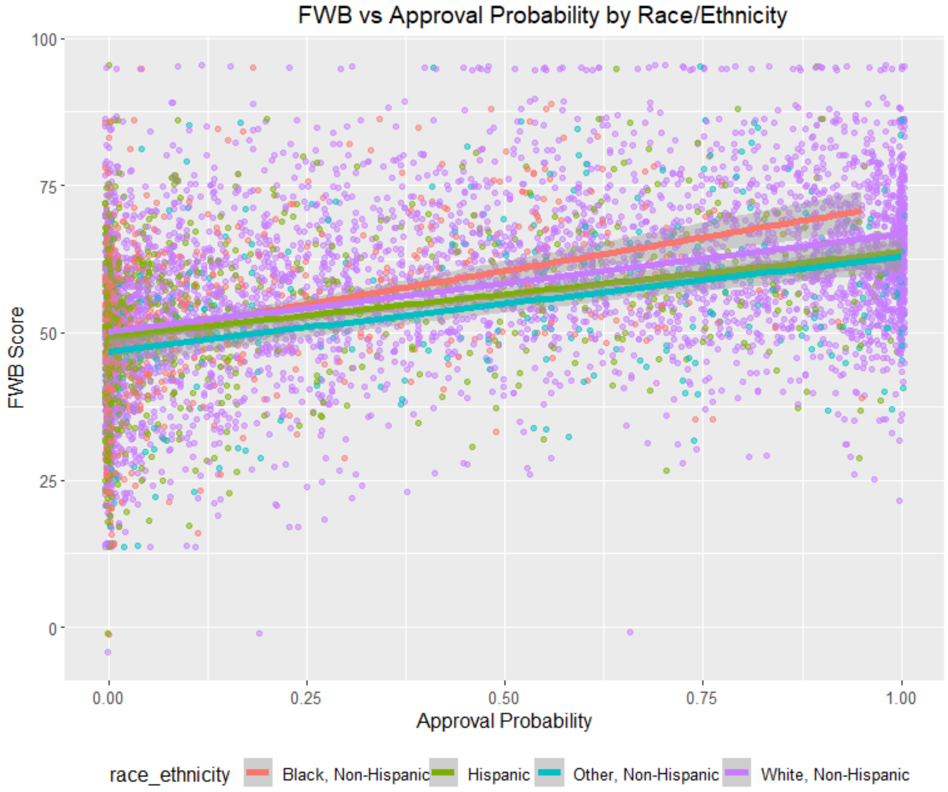
T-test comparing mean FWB score between those who have credit and those who do not

Mean, have credit	Mean, no credit	t	df	p	95% CI
57.56	44.08	25.31	6392	<2.2e-16	(12.43, 14.52)

T-test Results

Appendix 9

Relationship Between FWB and Approval Probability by Race/Ethnicity



Scatter plot of financial well-being score versus approval probability by race/ethnicity

	Black, Non-Hispanic	White, Non-Hispanic	Hispanic	Other, Non-Hispanic
Intercept	49.18 (<2e-16)	50.14 (<2e-16)	49.30 (<2e-16)	46.89 (<2e-16)
Prob_approved	22.68 (<2e-16)	16.55 (<2e-16)	14.42 (<2e-16)	16.04 (<2e-16)
R^2	0.147	0.195	0.100	0.155
Adj R^2	0.146	0.197	0.099	0.153
F-Stat	117.8	1088	96.56	61.42

Regression results of FWB score on approval probability for each race/ethnicity subset

Appendix 10

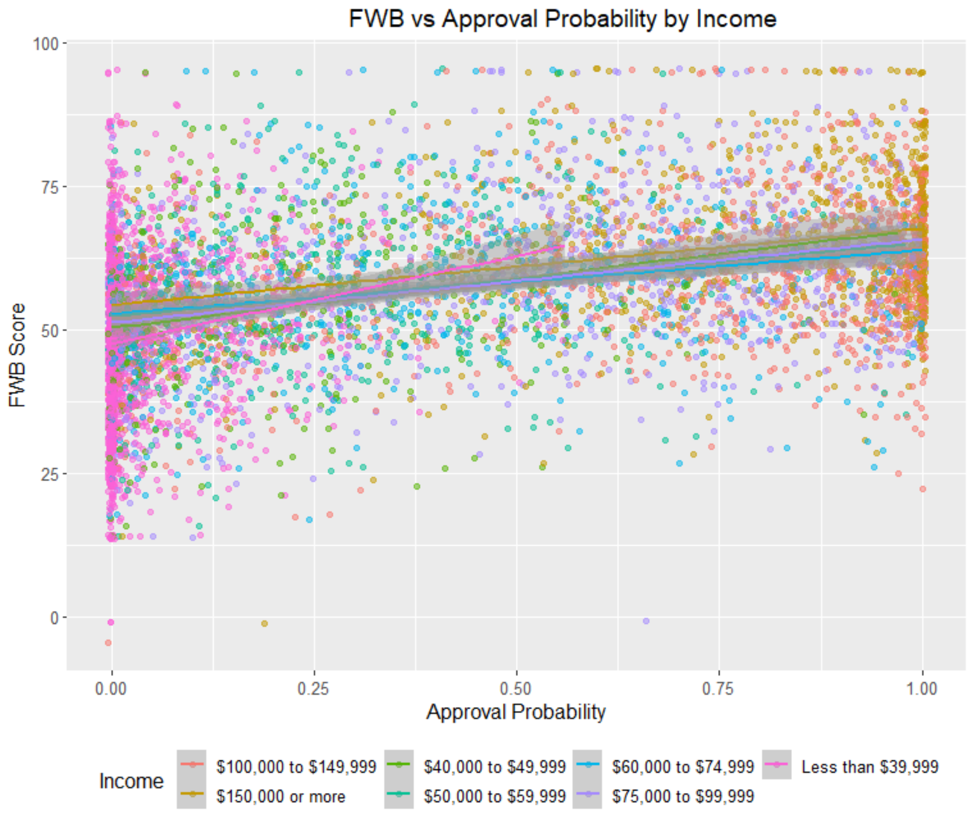
Mean Approval Probabilities and FWB Scores by Demographics

Demographic	Value	Mean Approval Probability	Mean FWB Score
Race/Ethnicity	Other, Non-Hispanic	0.472	54.46
Race/Ethnicity	White, Non-Hispanic	0.438	57.38
Race/Ethnicity	Hispanic	0.199	52.16
Race/Ethnicity	Black, Non-Hispanic	0.165	52.93
Income	$150,000 or more	0.739	64.15
Income	$100,000 to $149,999	0.671	60.15
Income	$75,000 to $99,999	0.511	58.63
Income	$60,000 to $74,999	0.365	56.76
Income	$50,000 to $59,999	0.265	55.70
Income	$40,000 to $49,999	0.202	53.77
Income	Less than $39,999	0.041	48.80
Savings Amount	$75,000 or more	0.695	68.94
Savings Amount	$20,000-74,999	0.680	63.80
Savings Amount	$5,000-19,999	0.504	59.14
Savings Amount	Unknown	0.387	57.31
Savings Amount	$0-4,999	0.103	46.92

Mean Approval Probabilities and FWB Scores Across Different Demographics

Appendix 11

Relationship Between FWB score and Approval Probability by Income



Scatter plot of financial well-being score versus approval probability by income

	Less than $39,999	$150,000 or more
Intercept	47.53 (<2e-16)	54.29 (<2e-16)
Prob_approved	30.63 (<2e-16)	13.34 (<2e-16)
R^2	0.030	0.099
Adj R^2	0.030	0.098
F-Stat	57.38	94.43

Regression results of FWB score on approval probability for highest and lowest income groups

Appendix 12

R Code

1_Data_Management.R

Perform data management tasks on LAR data set

2_Model_Fitting.R

Fit naïve Bayes, classification tree, and random forest models and evaluate performance

3_Analysis.R

Perform data management tasks on FWBS data set and use RF model for analysis