**Trabalho de Conclusão de Curso**

PÓS-GRADUAÇÃO EM CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL

**ALUNO: Daniel Favero**

**ORIENTADOR: Nome do Orientador**



Sumario

´

[1. ABSTRACT 2](#_Toc120819803)

[2. INTRODUCTION 3](#_Toc120819804)

[3. TRABALHOS RELACIONAIS 6](#_Toc120819805)

[4. ANALYSIS METHODOLOGY 7](#_Toc120819806)

[5. OBTAINED RESULTS 14](#_Toc120819807)

[6. RESULT ANALYSIS 22](#_Toc120819808)

[7. CONCLUSIONS AND FUTURE WORK 24](#_Toc120819809)

**ARTIGO CIENTÍFICO**

**Ethics on Tech: Explainable AI towards to a more Humane World**

# ABSTRACT

Under works

THIS WORK IS STILL DRAFT 😊

Github address: <https://github.com/danielfavero/ethical_mortgage_loans>

Please use GitHub for latest TCC versions

# INTRODUCTION

Discoursing previously on technology I have said that it “has become a center piece of the world. It is so ingrained in society that, for most of us, it is the first thing we interact with in the morning and the last thing we part from at night. This is, by and large, because technology has proven to improve our lives in innumerable ways” (Ethics on Tech p.3). It was presented then that the effect of technology on economy is uber evident. An estimate made by the Maddison project revealed that technology, seen after the 1st industrial revolution, boosted the global wealth by a factor of 50. The compound annual growth rate (CAGR) in the computer age shows about 75% faster growth in comparison to the advancements since the invention of the steam engine[[1]](#footnote-1). Erik Brynjolfsson, Senior Fellow at the Stanford Institute for Human-Centered AI, argues in his book The Second Machine Age that the computer age bent the social development curve. Erik concludes that “we’re living in a time of astonishing progress with digital technologies”. His foundation is sustained via “exponential improvement in […] computing, extraordinarily large amounts of digitized information, and recombinant innovation”. Brynjolfsson believes this to have been the beginning of a “second machine age [with the ability …] to create […] the emergence of real, useful artificial intelligence (AI) and the connection of most of the people on the planet via a common digital network” (Brynjolfsson, 2010, p. 53). PricewaterhouseCoopers, considered the second-largest professional services network in the world, published, in 2018, a report to “provide a clearer picture of the full economic potential of AI”. Applying analytical model analysis from 2017 to 2030, PwC concluded “that [the] global GDP could be up to 14% higher than this figure in 2030 as a result of AI – the equivalent of up to $15.7 trillion”. PWC argued that AI will augment the workforce and productivity”. (GILLHAM et al., 2018, p.6).

Clearly there seem to be a lot at stake, a few years ago, Google CEO Sundar Pichai said during a townhall event that “AI is one of the most important things humanity is working on. It is more profound than, I dunno, electricity or fire.” (CNBC). On top of it, the famous Andrew Ng, founder of the Google Brain Deep Learning Project and co-founder of Coursera said that “AI is the new electricity.”

With such high transformational expectations, the level of scrutiny on the technology can only grow larger. There are a number of examples showing where AI missed the mark. As part of a previous research, I have talked about bias and Inequality when people of color pay more for car insurance companies, fairness in an HR selection process, transparency as in the scandal that rocked Facebook with Cambridge Analytica development of psychometric models that were argued to have had an impact to the US election and privacy when argued that Pokémon, Go is a “carefully crafted and profitable algorithm used for private personal behavioral manipulation” (Ethics on Tech, p.16-17).

This work is going to dive deep into the Mortgage sector and how it is affecting minority classes to obtain mortgage loans. To give it some perspective. In 2021, analysis done by Black Knight, a provider of integrated software, data and analytics solutions in this area published that “mortgage originations hit a record high of $4.4 trillion” dollars (black knight, p.14). Another datapoint, shared by Federal Reserve Bank of New York, sums the Residential mortgage debt in the U.S. to $11.18 trillion as of the first quarter of 2022. This is clearly a relevant subject to maintain ethical standards and undergo severe scrutiny for bias.

In early 2017, the Pew Research Center published a study concluding that blacks and Hispanics face extra challenges in getting home loans. “In 2015, 27.4% of black applicants and 19.2% of Hispanic applicants were denied mortgages, compared with about 11% of white and Asian applicants, according to … data gathered under the federal Home Mortgage Disclosure Act” (Pew research). This is only one research, in May 2018, the Washington Post published a different study with 2017 data highlighting that “The overall rate of denials of mortgage applications from blacks was 18.4 percent …, with 13.5 percent for Hispanics and 10.6 percent for Asians. For non-Hispanic whites, it was 8.8 percent” (Washington Post). In 2021, “The Markup” published results from their analysis of 2019 data indicating that they had “found that lenders were 40 percent more likely to turn down Latino applicants for loans, 50 percent more likely to deny Asian/Pacific Islander applicants, and 70 percent more likely to deny Native American applicants than similar White applicants. Lenders were 80 percent more likely to reject Black applicants than similar White applicants.”

The goal of this project is to investigate whether machine learning explanability and interpretability can be used as tools to spot modeling biases in the mortgage industry by using the 2021 data from the Home Mortgage Disclosure Act.

# RELATED WORK

This is perhaps covered in other sections

# ANALYSIS METHODOLOGY

As part of our ETL (extract, transform, load) and for the purpose of this analysis, data on mortgage applications were used with the goal of creating several models that can predict a loan outcome (approved or denied). Data on mortgages are reported by the Federal Financial Institutions Examination Council via the Home Mortgage Disclosure Act – HMDA for short. The FFIE requires many financial institutions to maintain, report, and publicly disclose information about mortgages. As per the Consumer Financial Protection Act and Bureau, the:

“HMDA grew out of public concern over credit shortages [and] … Congress believed that some financial institutions had contributed to the decline of some geographic areas by their failure to provide adequate home financing to qualified applicants on reasonable terms and conditions. Thus, [showing the public] … whether financial institutions are serving the housing credit needs of the .. communities” (CFPB, p. 1).”

In this study, the 2021 HMDA data was used and is available for download at the CFPB website. The data schema has 85 fields and approximately 26million rows.

After data extraction, the first step in the data transformation was to execute data exploration. We chose the pandas-profiling library available at GitHub to help on this task. The pandas-profiling is an easy way to generate profile reports from a pandas DataFrame. While these reports can be very comprehensive, the minimal=True was needed to allow the system not to crash due to lack of memory. Combining pandas-profiling with an overall examination of the distribution loan outcomes per column, it was decided that 57 columns in the dataset were not needed. The determination was defined based on:

1. Columns that were added by the CFPB and were not needed;
2. Columns for observed race, ethnicity and sex were removed as to use reported data only;
3. Dropped columns with constant values like activity\_year, applicant\_ethnicity\_5;
4. Kept census\_tract but dropped derived\_msa\_md, state\_code and county\_code;
5. Dropped columns when their value is concentrated into a single category for more than 96% of the time;
6. Dropping columns that are not very representative in the denied loan outcome. If a column has less than 5% of loans that were denied, this column was excluded as it may not really contribute
7. Dropping columns with more than 70% missing values

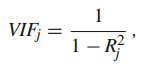
The next step in the data transformation is cleaning the features and target. The dataset was limited to loans that were either originated or denied, excluding all other loan outcomes (i.e. loan applications that were approved but not accepted, withdrawn by applicant, File closed for incompleteness, etc). At the same time, only conventional loans that were used for home purchases were used. This excluded loans guaranteed by veteran affairs, the federal housing administration or the USDA Rural Housing Service or Farm Service Agency. In addition, loans applied for home improvement, refinancing, cash-out refinancing or other purposes were also excluded. This seem a common practice in many of the analysis used to investigate loan biases.

Back to data transformation and to understand people’s origin, the attention was turned into race and ethnicity. Since these are reported in separate columns, it is necessary to combine them into a single categorical feature. The logic must be equally applied to both applicant and co-applicant. Next it was necessary to clean and simplify categorical features like Credit Models, Conforming Loan limits and to clean continuous variable like Debt-to-Income Ratio, Combine Loan-to-Value Ratio, Property value, loan Term and age.

This first transformation is then completed, and the HMDA dataset file can be exported for long term storage. The new generated file in then loaded in the notebook “2\_model\_creation.ipynb”. Note the new dataset contains continuous and categorical features. To minimize memory consumption, properly setting the feature dtypes was needed.

While performing the EDA (Exploratory Data Analysis), it was found that there were outliers. Those were identified via an investigation of each concerning features. Outliers were then converted into NA values. For dealing with such NA values, different mechanisms to fill them were investigated. Both Multivariate Imputation (MICE) and Fast KNN were investigated, however, given the size of the dataset, MICE didn't converge and Fast KNN was taking too long to process. It was decided that given the number of NAs to simply dropping these lines confident they were not detrimental to the results.

Next step was to investigate whether multicollinearity existed. According to Michael Allen’s Understanding Regression Analysis, “multicollinearity exists whenever an independent variable is highly correlated with one or more of the other independent variables in a multiple regression equation. Multicollinearity is a problem because it undermines the statistical significance of an independent variable. Other things being equal, the larger the standard error of a regression coefficient, the less likely it is that this coefficient will be statistically significant” (p.176). In the Concise Encyclopedia of Statistics by Yadolah Dodge, “A collinear relation between more than two variables will not always be the result of observing the pairwise correlations between the variables. A better indication of the presence of a collinearity problem is provided by variance inflation factors, VIF defined by:

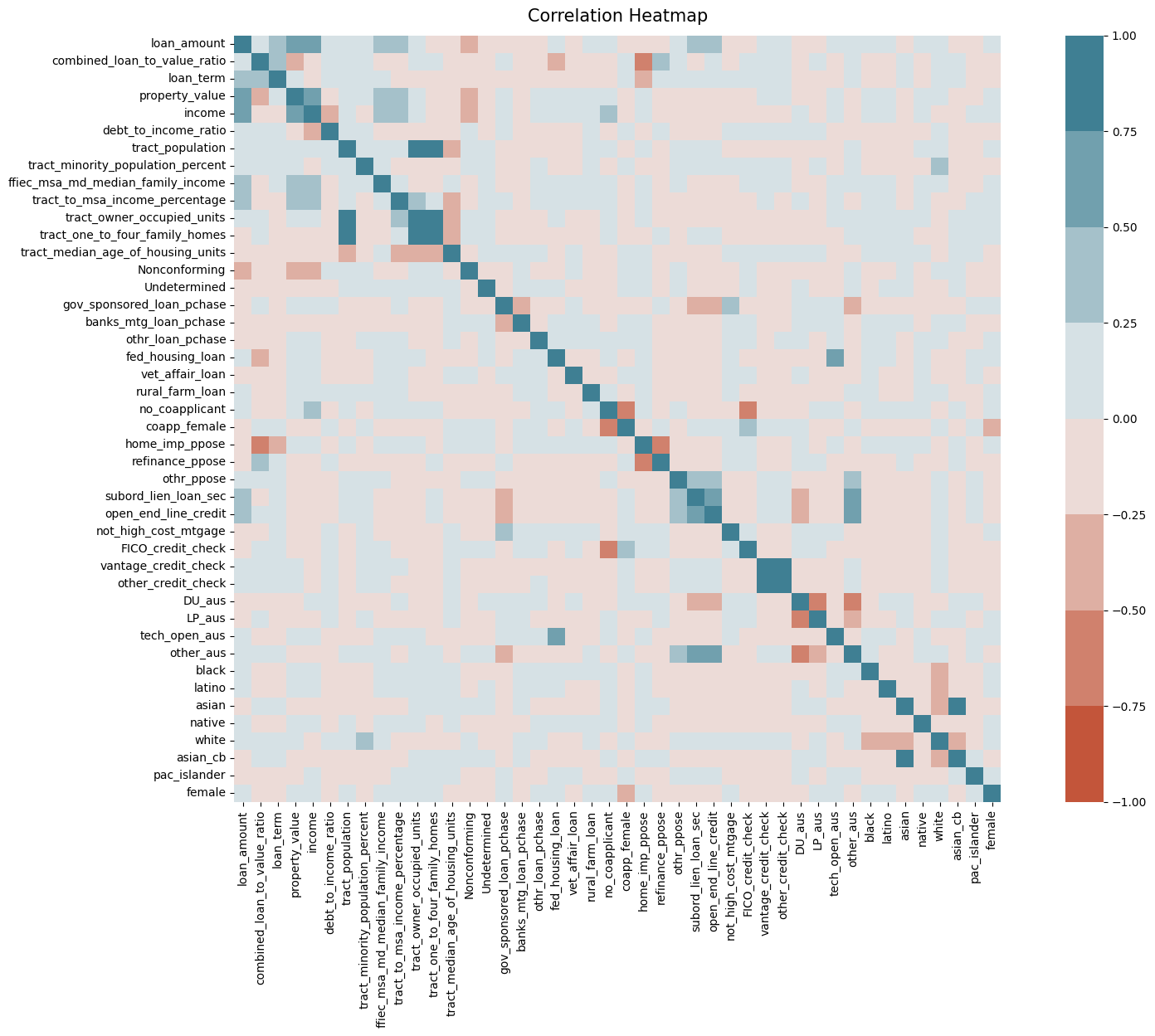
,

where R2, j is the coefficient of determination of the model.

To make GVIFs comparable across dimensions, (Fox and Monette 1992) also suggested reducing the GVIF to a linear measure, which is analogous to taking the square root of the usual VIF.

Since there are both continuous and categorical features in our dataset, Andrius Buteikis indicates in his “Practical Econometrics and Data Science” that for the continuous variables GVIF is the same as the VIF values before and for the categorical variables, we now get one GVIF value for each separate category type (session 4.5.3.2);

An analysis of both correlation matrix and the VIF indicated that removing asian\_cb, other\_aus, 'tract\_owner\_occupied\_units' fixed any concerning collinearity issues.



|  |  |  |  |
| --- | --- | --- | --- |
| *Features* | GVIF | VIF | Tolerance |
| loan\_amount | 2.512579 | 6.313053 | 0.158402 |
| property\_value | 2.347372 | 5.510156 | 0.181483 |
| tract\_population | 2.198557 | 4.833651 | 0.206883 |
| tract\_one\_to\_four\_family\_homes | 2.149277 | 4.619391 | 0.216479 |
| vantage\_credit\_check | 1.958439 | 3.835484 | 0.260723 |
| other\_credit\_check | 1.956557 | 3.828117 | 0.261225 |
| DU\_aus | 1.915598 | 3.669515 | 0.272516 |
| LP\_aus | 1.762323 | 3.105784 | 0.321980 |
| subord\_lien\_loan\_sec | 1.745398 | 3.046414 | 0.328255 |
| open\_end\_line\_credit | 1.688890 | 2.852348 | 0.350588 |
| coapp\_female | 1.638836 | 2.685784 | 0.372331 |
| no\_coapplicant | 1.613753 | 2.604198 | 0.383995 |
| tech\_open\_aus | 1.571849 | 2.470709 | 0.404742 |
| combined\_loan\_to\_value\_ratio | 1.549729 | 2.401659 | 0.416379 |
| gov\_sponsored\_loan\_pchase | 1.502814 | 2.258451 | 0.442781 |

A next step in the data analysis was to tackle the loan outcome imbalance. In the dataset, there were 2,473,435 Approved Loans while 197,662 were loans that got denied. This indicates a 12.51:1 proportion. Again, it was investigated different metrics for balancing the dataset. Synthetic Minority Over-sampling Technique for Nominal and Continuous features, like SmoteNC was tested and while there is good confidence this function would be able balance the dataset, it was opted to undersample the majority class (loans approved) understanding that 200 thousand samples were enough for model training.

**Model training**

In order to evaluate the explainability of a variety of models, several “glass” and “black” box models were selected. A glass box model refers to a model in which its parameters are known to the user, allowing for a transparent investigation of predictions. A black box model is one which its predictions cannot be easily explained. In addition, an inherent explainable model was selected. It is a model published from Microsoft called Explainable boosting machine.

As a summary, the models under investigation are:

* Glassbox models:
  + Decision Tree
  + Logistic Regression
* Blackbox modes:
  + Random Forest
  + Multilayer perceptron (MLP)
  + Support vector machine
* Explainable models:
  + Explainable Boosting Machine

While three blackbox models are trained only the two better ones will be used in the interpretation of final results.

In order to best optimize the blackbox models while controlling the computational budget, it was decided to use the RandomSearchCV algorithm. The resulting best parameters are in Tab.1. Note that a severely reduction to the dataset was applied during optimization. Even with such reduction, to run 20 iterations across all models with 5 fold cross validation, it took approximately 8 hours. Note that EBM and the Glassbox models were trained with default parameters.

Table 1 Best params for Black box models

|  |  |  |
| --- | --- | --- |
| model | best\_score | best\_params |
| MLP | 0.508953 | {'verbose': False, 'max\_iter': 1000, 'learning\_rate\_init': 3.727593720314938e-05, 'hidden\_layer\_sizes': (1500, 3), 'alpha': 0.001} |
| RandForrest | 0.504899 | {'n\_estimators': 200, 'min\_samples\_split': 8, 'min\_samples\_leaf': 2, 'max\_depth': 80, 'bootstrap': True} |
| svm | 0.503547 | {'kernel': 'linear', 'C': 10} |

Once models were trained, each model was explored for explainability.

For the decision tree classifier, the Scikit .feature\_importances\_ was used. It works based on a score that estimates how significant they are at predicting a target variable. The scores are calculated on the weighted Gini indices. The data scientist Stacey Ronaghan, in a 2018 post at towardsdatascience explained that “feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node”. She continues that “Scikit-learn calculates a nodes importance using Gini Importance, assuming only two child nodes”.

Another simple explainable glass box model used was the logistics regression. Essentially, the model coefficients are the log of the odds for each feature. For interpretability, the odds ratio can be calculated by taking the exponential of the model coefficients.

In 2019, researchers at Microsoft published a python open-source library called InterpretML, with the goal of opening up tools for model/algorithm explainability. As the paper elaborates:

“InterpretML exposes two types of interpretability – glassbox, which are machine learning models designed for interpretability (ex: linear models, rule lists, generalized additive models), and blackbox explainability techniques for explaining existing systems (ex: Partial Dependence, LIME). The package enables practitioners to easily compare interpretability algorithms by exposing multiple methods under a unified API, and by having a built-in, extensible visualization platform” (InterpretML p.1).

Important for the use in this work, the InterpretML library contains the implementation of the Explainable Boosting Machine (EBM). Despite being a glassbox model it claims to bring as high accuracy as blackbox models. EBM uses modern machine learning techniques like bagging and boosting to breathe new life into traditional GAMs (Generalized Additive Models). This makes them as accurate as random forests and gradient boosted trees, and also enhances their intelligibility and editability (InterpretML p.3).

Other explainable techniques were available before. In 2016, Marco Ribeiro along with Sameer Singh and Carlos Guestrin from the University of Washington proposed the Local Interpretable Model-Agnostic Explanations algorithm- LIME for short. As it claims, LIME can “explain the predictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model” (Why should I trust you, p.1). Dr. Ribeiro wanted to build interpretability algorithm that was agnostic to any model. He explains this can be achieved by making perturbation in the input while monitoring changes to the model predictions. This turns out to be a benefit in terms of interpretability, because we can perturb the input by changing components that make sense to humans (e.g., words or parts of an image), even if the model is using much more complicated components as features (e.g., word embeddings).

In 2017, Scott Lundberg and Su-In Lee, also from the University of Washington and current researchers at Microsoft published yet another approach to interpreting model predictions. Their research is grounded on the fact that the *explanation* for a model’s prediction is as crucial and the model’s accuracy itself. In order to ease the model interpretability for researchers and practitioners, they presented a “unified framework” called SHAP (SHapley Additive exPlanations). SHAP works by assigning “each feature an importance value for a particular prediction”. It works based on the “identification of a new class of additive feature importance measures and results showing there is a unique solution in this class with a set of desirable properties” (SHAP p.1).

To further elaborate, Sharayu Rane at TDS breaks it down to say that:

“Shapely values are obtained by incorporating concepts from Cooperative Game Theory and local explanations. Given a set of players, Cooperative Game Theory defines how well and fairly to distribute the payoff amongst all the payers that are working in coordination…each feature represent the change in the expected model prediction when conditioning on that feature, [explaining] the [contributing].. difference between the average model prediction and the actual prediction of the instance” (Rane, 2019).

# OBTAINED RESULTS

Results from the model training are evaluated using F1 accuracy score. Despite the overall accuracy, F1 scores for the individual classes were evaluated. Class 0 means denied loans while class 1 means approved loans.

Table 2 Model Result Accuracy

|  |  |  |  |
| --- | --- | --- | --- |
| model | f1-score\_0 | f1-score\_1 | f1-score\_accuracy |
| EBM | 89.83% | 87.91% | 88.95% |
| Random Forest | 89.83% | 87.80% | 88.91% |
| Logistic Regression | 89.35% | 86.90% | 88.25% |
| MLP | 88.06% | 87.12% | 87.61% |
| Decision Tree | 82.02% | 75.49% | 79.26% |

**Decision Tree results**

Figure 1 uses the Scikit-learn feature importance to indicate the features that brought the highest weighted Gini indices.

Chart

Description automatically generated

Figure Decision Tree Features by importance

**Logistics Regression results**

Detailed Logit results were obtained using the .summary().

|  |  |  |  |
| --- | --- | --- | --- |
| Logit Regression Results | | | |
| Dep. Variable: | loan\_outcome | No. Observations: | 395324 |
| Model: | Logit | Df Residuals: | 395283 |
| Method: | MLE | Df Model: | 40 |
| Date: | Wed, 30 Nov 2022 | Pseudo R-squ.: | 0.6187 |
| Time: | 20:51:20 | Log-Likelihood: | -1.0447e+05 |
| converged: | True | LL-Null: | -2.7402e+05 |
| Covariance Type: | nonrobust | LLR p-value: | 0.000 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | coef | std err | z | P>|z| | [0.025 | 0.975] |
| loan\_amount | 0.0322 | 0.014 | 2.300 | 0.021 | 0.005 | 0.060 |
| combined\_loan\_to\_value\_ratio | -0.0728 | 0.006 | -11.747 | 0.000 | -0.085 | -0.061 |
| loan\_term | -0.0396 | 0.005 | -7.389 | 0.000 | -0.050 | -0.029 |
| property\_value | 0.0152 | 0.013 | 1.144 | 0.253 | -0.011 | 0.041 |
| income | -0.5825 | 0.293 | -1.987 | 0.047 | -1.157 | -0.008 |
| debt\_to\_income\_ratio | -0.6513 | 0.006 | -111.746 | 0.000 | -0.663 | -0.640 |
| tract\_population | 0.0197 | 0.011 | 1.825 | 0.068 | -0.001 | 0.041 |
| tract\_minority\_population\_percent | -0.0469 | 0.007 | -6.599 | 0.000 | -0.061 | -0.033 |
| ffiec\_msa\_md\_median\_family\_income | 0.2462 | 0.007 | 37.450 | 0.000 | 0.233 | 0.259 |
| tract\_to\_msa\_income\_percentage | 0.1117 | 0.006 | 18.482 | 0.000 | 0.100 | 0.124 |
| tract\_one\_to\_four\_family\_homes | 0.0670 | 0.010 | 6.503 | 0.000 | 0.047 | 0.087 |
| tract\_median\_age\_of\_housing\_units | 0.1055 | 0.006 | 17.036 | 0.000 | 0.093 | 0.118 |
| Nonconforming | -0.7037 | 0.022 | -32.107 | 0.000 | -0.747 | -0.661 |
| Undetermined | 23.8715 | 0.781 | 30.565 | 0.000 | 22.341 | 25.402 |
| gov\_sponsored\_loan\_pchase | -10.4455 | 0.277 | -37.643 | 0.000 | -10.989 | -9.902 |
| banks\_mtg\_loan\_pchase | -10.3946 | 0.447 | -23.237 | 0.000 | -11.271 | -9.518 |
| othr\_loan\_pchase | -8.2362 | 0.289 | -28.491 | 0.000 | -8.803 | -7.670 |
| coapp\_white | -0.1664 | 0.021 | -8.088 | 0.000 | -0.207 | -0.126 |
| coapp\_black | 0.1044 | 0.049 | 2.135 | 0.033 | 0.009 | 0.200 |
| coapp\_latino | 0.1062 | 0.033 | 3.219 | 0.001 | 0.042 | 0.171 |
| coapp\_asian | 0.1200 | 0.033 | 3.661 | 0.000 | 0.056 | 0.184 |
| coapp\_native | 0.2784 | 0.133 | 2.100 | 0.036 | 0.019 | 0.538 |
| coapp\_pac\_islander | 0.6185 | 0.230 | 2.685 | 0.007 | 0.167 | 1.070 |
| coapp\_female | -0.2113 | 0.021 | -10.080 | 0.000 | -0.252 | -0.170 |
| subord\_lien\_loan\_sec | -0.1630 | 0.034 | -4.817 | 0.000 | -0.229 | -0.097 |
| open\_end\_line\_credit | 0.5878 | 0.036 | 16.458 | 0.000 | 0.518 | 0.658 |
| FICO\_credit\_check | -0.3723 | 0.020 | -18.290 | 0.000 | -0.412 | -0.332 |
| vantage\_credit\_check | 0.1330 | 0.046 | 2.881 | 0.004 | 0.043 | 0.224 |
| other\_credit\_check | -0.4078 | 0.039 | -10.476 | 0.000 | -0.484 | -0.331 |
| coapp\_FICO\_credit\_check | -0.3358 | 0.017 | -19.489 | 0.000 | -0.370 | -0.302 |
| coapp\_vantage\_credit\_check | -0.1358 | 0.595 | -0.228 | 0.819 | -1.301 | 1.030 |
| coapp\_other\_credit\_check | 0.8269 | 0.489 | 1.692 | 0.091 | -0.131 | 1.785 |
| DU\_aus | 0.0388 | 0.013 | 2.905 | 0.004 | 0.013 | 0.065 |
| LP\_aus | -0.0759 | 0.018 | -4.330 | 0.000 | -0.110 | -0.042 |
| tech\_open\_aus | 0.8987 | 0.194 | 4.641 | 0.000 | 0.519 | 1.278 |
| black | 0.5040 | 0.026 | 19.339 | 0.000 | 0.453 | 0.555 |
| latino | 0.3194 | 0.020 | 15.748 | 0.000 | 0.280 | 0.359 |
| asian | 0.1181 | 0.021 | 5.711 | 0.000 | 0.078 | 0.159 |
| native | 0.6116 | 0.081 | 7.585 | 0.000 | 0.454 | 0.770 |
| pac\_islander | 0.6077 | 0.157 | 3.859 | 0.000 | 0.299 | 0.916 |
| female | -0.0936 | 0.014 | -6.757 | 0.000 | -0.121 | -0.066 |

**Explainable Boosting Machine results**

By using the model\_EB.explain\_global(), an overall mean absolute score ranked by feature is plotted as observed in Figure2.

Chart

Description automatically generated

Figure Explainable Boosting Machine Mean Absolute Score. Overall Importance

Below Fig.3 is the model explanation for debt-to-income ratio (DTI). To understand the range, Fig.4 indicates the relationship between DTI and the standardized DTI used to train the model.

A picture containing chart

Description automatically generated

Figure 3 Global Explainer: Debt-to-Income ratio

Chart, scatter chart

Description automatically generated

Figure 4 Map of original DTI vs Standardized DTI used to feed the models

To understand the direct impact of race into the model, global explanation was plotted for both Black and Latino individuals.

A picture containing chart

Description automatically generated

Figure EBM feature weight for black individuals

Timeline

Description automatically generated with low confidence

Figure EBM feature weight for latino individuals

In order to dig deeper into the model explanations and to understand which features were considered important to determine the model predictions for loan denials. It was used the explain\_local() function for one sample of each of the 4 most biased races: Latinos, Blacks, Native Americans and Pacific Islanders. The same method was used for the EBM, MLP and Random Forrest models.

Sample results for EBM are presented below:

Sample: Latino Individual

Chart

Description automatically generated

Figure EBM explain local of denied loan for a Latino individual

Sample: Black Individual

Chart, bar chart

Description automatically generated

Figure EBM explain local of denied loan for a black individual

Sample: Native American

Chart, funnel chart

Description automatically generated

Figure EBM explain local of denied loan for a native American individual

Sample Pacific Islander

Chart

Description automatically generated

Figure EBM explain local of denied loan for a Pacific Islander individual

**Multi-layer perceptron results with LIME**

Sample: Latino Individual

Chart, bar chart, funnel chart

Description automatically generated

Figure LIME explain local of denied loan for a latino individual

Sample: Black Individual

Chart, bar chart

Description automatically generated

Figure LIME explain local of denied loan for a black individual

Sample: Native American

Chart, bar chart

Description automatically generated

Figure LIME explain local of denied loan for a native American individual

Sample: Pacific Islander

Chart, bar chart

Description automatically generated

Figure LIME explain local of denied loan for a Pacific Islander individual

**Random Forrest results with SHAP**

From SHAP values, the shap.summary\_plot() can be used to provide a ranking, based on the average feature impact on the model output as in Figure13.

Chart

Description automatically generated

Figure Average Feature impact for SHAP

As it was done for EBM and Lime, below are the sample results from SHAP for the Random Forrest Model.

Sample: Latino Individual

Chart

Description automatically generated

Figure SHAP explain local of denied loan for a Latino individual

Sample: Black Individual

Chart

Description automatically generated

Figure SHAP explain local of denied loan for a black individual

Sample Native American

Chart, waterfall chart

Description automatically generated

Figure SHAP explain local of denied loan for a native American individual

Sample Pacific Islander

Chart

Description automatically generated

Figure SHAP explain local of denied loan for a Pacific Islander individual

# RESULT ANALYSIS

**Interpreting the Decision Tree results**

The overall model F1 Score Accuracy was on 79.26% with an 82% accuracy for the Class 0 (loan denied). Investigating the DT feature importance, it can be concluded that 90% of the model can be explained by 4 main features. Government sponsored loan purchases indicate whether Fannie Mae, Ginnie Mae, Freddie Mac or Farmer Mac are purchasing the loan from the lending institution. The Debt-to-Income ratio is a clear measure of an applicant’s ability to pay the mortgage. Bank loan purchase indicate whether a Commercial bank, savings bank, or savings association is purchasing the loan from the lender. Finally, Home purchase indicates whether the applicant is intending to purchase a home instead of seeking to refinance or to make improvements to the home. Not much importance can be extracted from an individual race by using a decision Tree Classifier.

**Interpreting the Logistics Regression results**

The model performance was surprisingly better than the MLP, at an overall F1 score of 88.3% and 89% for Class 0. As elaborated in section 4, we can obtain the relation between variables via observing the exponential of the Logit model coefficients. By filtering only, the race/ethnic related variables, Blacks are 66% more likely to be denied a loan compared to non-hispanic whites. Latinos are 38% more likely to be denied while Native Americans are the worse with 84% more chance of denial.

|  |  |
| --- | --- |
| features | odds\_ratio |
| native American | 1.843340 |
| Pacific Islander | 1.836214 |
| Black | 1.655370 |
| latino | 1.376250 |
| asian | 1.125397 |

These results appear aligned with the ones published by the 2019 Markup article as mentioned earlier. As for another datapoint. In September 2022, the CFPB put out a report of the 2021 Mortgage trends. In this report, they indicated that:

“In past years, Black and Hispanic White borrowers had notably higher denial rates in 2021 than non-Hispanic White and Asian borrowers. Among home purchase applications, the denial rates were 15.3 percent for Black applicants and 10.6 percent for Hispanic[s]... In contrast, the denial rates of home purchase applications were 7.9 percent for Asian applicants and 6.3 percent for non-Hispanic White applicants” (CFPB. p.29).

CFPB statement indicates an even worse picture as observed by the Logit results above.

**Interpreting the other models along with LIME, SHAP and EBM**

It is evident from the Logit results as well as the several reports published by the media outlets that loan denials are disproportionally high for minority races. The MLP delivered a performance only batter than a decision tree classifier. It brought an 87.6% overall F1 Score accuracy. Using LIME, it was possible to peak under the hood as to what features were most important. Using a function to explain local individual samples, it is observed that both Income and loans purchased government-sponsored financial companies were the most important features (see Figures 9-12). Debt-to-income ratio is not listed as an important feature for the model. In any of the samples, race appeared as a factor in the model’s loan classification.

The MLP results are dissonant when contrasted against the Random Forest classifier investigated using SHAP. Firstly, the Random Forrest classifier showed a 130 basis points better overall accuracy (88.91%) compared to the MLP. The overall mean SHAP values indicate that Gov. sponsored loan purchases and income are most important features followed by Debt-to-income ratio. The mean SHAP value indicate that minority races have a small impact to the model’s classification. For the individual local predictions, SHAP indicates in the Latino Sample (Fig. 14) that being a Latino affected 2.7% of the model’s decision. The Black sample indicates that being black affects negatively the prediction by 2.1%. (Fig. 15). Race was not a relevant factor in neither the Native American nor the Pacific Islandic samples.

InterpretML paper bragged of the EBM’s accuracy, and it delivered! The model is the most accurate in this exercise. With 89% overall accuracy, it nearly reached 90% F1 score for Class 0 (Loan denied). The model explanations also appear the most relevant. It puts the two most important features as who purchases the loan from the lender (government sponsored or banks) and as a third feature, debt-to-income ratio. InterpretML global explainer can also tell that a Debt-to-Income ratio higher than 50% has a detrimental impact to someone’s ability to seek a loan approval (Fig. 3, 4)

When looking at the explain local results, where individual predictions were investigated. The explanations using EBM did highlight that races are within the top 5-10 most features responsible for a loan denial (Figg 7-10). For more specifics, Fig. 8 shows the sample for the Black individual. In this sample, it can be seen that being Black was the fifth most relevant feature contributing to the model’s denial prediction.

# CONCLUSIONS AND FUTURE WORK

For all models and metrics investigated, all of them, but the Multi-layer perceptron explained via LIME, showed that race impacted the model’s classification for a denied loan. Models like EBM indicated race had a bigger role in the model’s decision compared to Random Forest with SHAP. The logistics regression made a clear estimate for how much worse minority races, like Hispanic, Black, native American and Pacific Islanders, getting denied in comparison with non-Hispanic whites.

These outcomes were already expected. The overwhelming amount of media coverage stressing these biases strengthens this report’s findings that minority races are more likely to have a loan denied. But why is that? Is there hope? Can Explainable AI help opening trenches to more ethical behaviors?

Lenders and banks have repeatedly assured the public that there are no biases in their loan models. In response to “The Markup” article, Blair Bernstein, director at the ABA (American Bankers Association), acknowledged that mortgage metrics shows disparities but unfortunately there is a “given limitations” to the public data available via HDMA, “the numbers are not sufficient on their own to explain why those disparities exist” (The Markup). He added that the “ABA firmly believes that discrimination has no place in the mortgage market. It is not just morally wrong but a violation of federal and state law”. The conclusion from ABA is that “individual factors such as a borrower’s credit score and credit history … can help explain why seemingly comparable applicants may not always end up with the same lending outcome.”

There are supposedly fairer credit models to help address some of the inconsistencies. A 2021 report by Vantagescore, indicated that their “models …provide a fair and accurate credit score for 37 million more consumers than the population scored by conventional models (like FICO), covering 96% of all U.S. consumers who are 18 or older” (vantage p.2). They claim that the secret sauce is in their state-of-the art machine learning modeling.

VantageScore is a new credit model – created by Experian, TransUnion and Equifax credit bureaus – to compete with FICO. The above claims were published as part of the release of Vantage Score 4.0, hoping to “help provide scores for consumers with minimal credit histories or thin credit files” (nerdwallet). The claim is that “VantageScore 4.0 and… VantageScore 3.0, are fundamentally similar” with the caveat of a “few new tweaks and characteristics” that would bring the ability “to score about 37 million more U.S. adults than other models”.

The HMDA data has no reporting data for Vantage Score 4.0 yet, but it does indicate when a Vantage Score 3.0 has been used. An analysis of the 2021 data for VantageScore 3.0 didn’t give a lot of hope yet. It worth caveating that VantageScore only promised that their latest model would allow an additional 37 million adults to be scored, but it did not say anything about being approved. In fact, data shows that non-Hispanic whites are approved for a conventional loan 47% of the time when the model used to generate the credit score was the VantageScore 3.0. This number drops to 32% and 21% for Latinos or Blacks. In contrast, when FICO score is used, non-Hispanic whites are approved 92% of the time while Latinos and Blacks are approved 85% and 77% of the time. These are simple quantifications and further analysis of other parameters – like debt-to-income ratio, loan amount, etc – could be used to better understand such discrepant metrics.

The fact is that VantageScore 4.0 is likely not going to eliminate any race disparities, reports indicate that these disparities come from long term discriminatory practices against minority classes. While this report is not going to investigate this further, it appears clear that the use of Explainable AI analysis can help us see such racial disparities but the answer to more ethical algorithms is in long term analysis impacts. VantageScore 4.0 is indeed a step in the right direction.

The role of the media is key in continued investigation of malpractices. For scoring companies, banks and lenders developing algorithms should always employ strict, prioritized practices to avoid biases. An ethical company will always require more cost and effort.

In the investigations of this report, it was observed that MLP with LIME had very different feature weight prediction in comparison to models like EBM and Random Forrest with SHAP. This could be an area of further investigation to understand how LIME would interpret the Random Forrest model or how SHAP would understand the MLP. Other more advanced algorithms were also not used and a CNN (convolutional Neuro Network) could be a next step.

WORKS CITED

A picture containing indoor, dark, night sky

Description automatically generatedAQUINAS, Thomas. **Summa Theologica:**Complete & Unabridged. Tradução: Fathers of the English Dominican Province. Kindle ed. USA: Coyote Canyon Press, 2010.

ARJO, Dennis; GALE, Dawn; CONRAD, Omar. **Ethics:**Introduced. 1 ed. USA : Cognella, 2019.

BARLETT, Robert C.; COLLINS, Susan D.. **Aristotle's Nicomachean Ethics.**Kindle ed. USA : University of Chicago Press, Aristotle. Aristotle's Nicomachean Ethics . University of Chicago Press., 2011.

Brynjolfsson, Erik, and Andrew McAfee. **The Second Machine Age**: Work, Progress, and Prosperity in a Time of Brilliant Technologies. New York: W.W. Norton & Company, 2014.

COMPANIESMARKETCAP.COM. **CompaniesMarketCap.com.**Largest Companies by Market Cap. *[S.l.].*., 2021. Disponível em: https://companiesmarketcap.com/. Acesso em: 18 jul. 2021.

ELON Musk, ISS R&D Conference, July 19, 2017. UK, 2017. 1 vídeo (1:27:11). Publicado pelo Space Policy and Politics. Disponível em: https://www.youtube.com/watch?v=BqvBhhTtUm4. Acesso em: 13 ago. 2021.

GILLHAM, Jonathan *et al*. . **The macroeconomic impact of artificial intelligence.**web. 2018. Disponível em: https://www.pwc.co.uk/economic-services/assets/macroeconomic-impact-of-ai-technical-report-feb-18.pdf. Acesso em: 14 jul. 2021.

HOW Cambridge Analytica Exploited the Facebook Data of Millions. USA, 2018. 1 vídeo (2:32min). Publicado pelo The New York Times. Disponível em: https://www.youtube.com/watch?v=mrnXv-g4yKU. Acesso em: 15 ago. 2021.

INVESTOR'S BUSINESS DAILY. **Investor's Business Daily.**Venture Capital Flooding Into Technology Companies At Record Levels. *[S.l.].*., 2021. Disponível em: https://www.investors.com/news/technology/venture-capital-flooding-into-tech-startups-at-record-levels/. Acesso em: 18 jul. 2021.

KANT, Immanuel. **Groundwork for the metaphysics of morals/ Kant, Immanuel. Groundwork for the Metaphysics of Morals :**Rethinking the Western tradition. Tradução: Frank M. Turner; Allen W. Wood. Kindle ed. USA: Yale University Press, 2002.

LIVING without modern technology. UK, 2009. 1 vídeo (4:21). Publicado pelo The Guardian. Disponível em: https://www.youtube.com/watch?v=llN03ApB07M. Acesso em: 13 ago. 2021.

MCCARTHY, John. WHAT IS ARTIFICIAL INTELLIGENCE?. *ln:*Stanford University. **..**Stanford, CA 94305, 24 nov. 2004. Disponível em: https://homes.di.unimi.it/borghese/Teaching/AdvancedIntelligentSystems/Old/IntelligentSystems\_2008\_2009/Old/IntelligentSystems\_2005\_2006/Documents/Symbolic/04\_McCarthy\_whatisai.pdf. Acesso em: 18 jul. 2021.

MILL, John Stuart. **Utilitarianism:**Illustrated. Classic Edition ed. Kindle online: Green World Publishing, 2016.

MORRIS, Ian. **Why the West rules:**for now. United States: Farrar, Straus and Giroux, 2010.

O'NEIL, Cathy. **Weapons of Math Destruction.**1 ed. USA: Broadway Books, 2016.

POJMAN, Louis. **Ethics:**Discovering Right and Wrong. 3. ed. USA: Wadsworth Publishing Company, 1990.

POJMAN, Louis and Fieser, James. **Ethics:**Discovering Right and Wrong. 8. ed. USA: Cengage Learning, 2015.

REAMER, Andrew. **The Impacts of Technological Invention on Economic Growth:**A Review of the Literature. 2014. The George Washington Institute of Public Policy, The George Washington University, USA, 2014. Disponível em: https://gwipp.gwu.edu/sites/g/files/zaxdzs2181/f/downloads/Reamer\_The\_Impacts\_of\_Invention\_on\_Economic\_Growth\_02-28-14.pdf. Acesso em: 11 jul. 2021.

SHARON A. LLOYD. **Stanford Encyclopedia of Philosophy.***[S.l.].*The Metaphysics Research Lab, 2018. Disponível em: https://plato.stanford.edu/entries/hobbes-moral/. Acesso em: 13 ago. 2021.

SINGER, Peter. **Practical Ethics.**2. ed. USA: Cambridge University Press, 1999.

UNIVERSITY OF GRONINGEN. **Maddison Project Database 2020.***[S.l.].*Groningen Growth and Development Centre, 2020. Disponível em: https://www.rug.nl/ggdc/historicaldevelopment/maddison/releases/maddison-project-database-2020. Acesso em: 14 jul. 2021.

ZUBOFF, Shoshana. **The Age of the Surveillance Capitalism:**The fight for a human future at the new frontier of power. 1 ed. USA: Public Affairs, 2019.

<https://www.nerdwallet.com/article/finance/vantagescore-4-0>

<https://vantagescore.com/credit-invisible-no-longer-examining-the-relationship-between-socioeconomic-disparities-and-scoreability/>

<https://www.blackknightinc.com/wp-content/uploads/2022/03/BKI_MM_Jan2022_Report.pdf>

<https://www2.deloitte.com/content/dam/Deloitte/us/Documents/deloitte-analytics/us-ai-institute-state-of-ai-fifth-edition.pdf>. Accessed 11/21

<https://www.pewresearch.org/fact-tank/2017/01/10/blacks-and-hispanics-face-extra-challenges-in-getting-home-loans/> Accessed 11/24

<http://pewrsr.ch/2iYR1hC>

<https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms> Accessed 11/24

Home Mortgage Disclosure Act

Examination Procedures

<https://www.cfpaguide.com/portalresource/manual%20-%20hmda.pdf>

<https://www.federalreserve.gov/boarddocs/caletters/2009/0910/09-10_attachment.pdf>

HMDA Data

<https://ffiec.cfpb.gov/data-publication/snapshot-national-loan-level-dataset/2021>

(1997). The problem of multicollinearity. In: Understanding Regression Analysis. Springer, Boston, MA. https://doi.org/10.1007/978-0-585-25657-3\_37

<https://doi.org/10.1007/978-0-585-25657-3_37>

Dodge, Y. (2008) Least Significant Difference Test. In: Dodge, Y., Ed., The Concise Encyclopedia of Statistics, Springer, New York.

<http://www.stewartschultz.com/statistics/books/The%20Concise%20Encyclopedia%20of%20Statistics.pdf>

PAGE 102

Practical Econometrics and Data Science

Andrius Buteikis

<http://web.vu.lt/mif/a.buteikis/wp-content/uploads/PE_Book/>

Fox, John, and Georges Monette. “Generalized Collinearity Diagnostics.” *Journal of the American Statistical Association*, vol. 87, no. 417, 1992, pp. 178–83. *JSTOR*, [https://doi.org/10.2307/2290467. Accessed 26 Nov. 2022](https://doi.org/10.2307/2290467.%20Accessed%2026%20Nov.%202022).

Mice

<https://www.jstatsoft.org/article/view/v045i03>

Fast KNN

https://impyute.readthedocs.io/en/master/\_modules/impyute/imputation/cs/fast\_knn.html

Stacey Ronaghan, The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark

<https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3>

smote NC

<https://www.jair.org/index.php/jair/article/view/10302>

<https://www.jair.org/index.php/jair/article/view/10302/24590>

@article{nori2019interpretml,

title={InterpretML: A Unified Framework for Machine Learning Interpretability}, author={Nori, Harsha and Jenkins, Samuel and Koch, Paul and Caruana, Rich},

journal={arXiv preprint arXiv:1909.09223},

year={2019}

}

"Why Should I Trust You?": Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin

<https://arxiv.org/abs/1602.04938>

A Unified Approach to Interpreting Model Predictions

Scott Lundberg, Su-In Lee

<https://arxiv.org/abs/1705.07874>

SHAP: A reliable way to analyze model interpretability

Sharayu Rane

<https://towardsdatascience.com/shap-a-reliable-way-to-analyze-your-model-interpretability-874294d30af6>

CONSUMER FINANCIAL PROTECTION BUREAU | SEPTEMBER 2022

Data Point: 2021 Mortgage Market Activity and Trends

<https://files.consumerfinance.gov/f/documents/cfpb_data-point-mortgage-market-activity-trends_report_2022-09.pdf>

1. The average global GDP per capita at the beginning of the first industrial revolution, in 1820, was $1,102. This number raised to $5,952 in the beginning of the computer age in 1970. In contrast, the last available GDP data, from 2018, is of $15,212. Calculating the GDP Compound annual growth rate (CAGR) pre and post 1970 would give us 1.13% (from 1820 to 1970) and 1.97% (from 1970 to 2018). Additionally, estimates made by Angus Maddison also reveal that the GDP CAGR from the beginning of the common era till 1820 was about 0.02% (University, 2020). [↑](#footnote-ref-1)