A Study of Adoption of Cloud computing and AI and ML in FinTech Sector

Sree Supraja Rama Sumitranandan Kamarajugadda

Final Thesis Report

August 2022

**Table of Contents**

[LIST OF EQUATIONS 4](#_Toc112710319)

[LIST OF TABLES 4](#_Toc112710320)

[LIST OF FIGURES 4](#_Toc112710321)

[1. INTRODUCTION 7](#_Toc112710322)

[1.1. ABSTRACT 7](#_Toc112710323)

[1.2. OVERVIEW OF CLOUD COMPUTING AND MACHINE LEARNING 8](#_Toc112710324)

[1.3. PROBLEM STATEMENT 10](#_Toc112710325)

[1.4. AIMS AND OBJECTIVES OF THE RESEARCH 11](#_Toc112710326)

[1.5. RESEARCH QUESTIONS 12](#_Toc112710327)

[1.6. SCOPE OF THE RESEARCH 12](#_Toc112710328)

[1.7. SIGNIFICANCE OF THE RESEARCH 12](#_Toc112710329)

[1.8. STRUCTURE OF THE RESEARCH 13](#_Toc112710330)

[2. LITERATURE REVIEW 14](#_Toc112710331)

[2.1. OVERVIEW OF FINTECH 14](#_Toc112710332)

[2.2. CLASSIFICATION OF SERVICES AND PRODUCTS IN FINTECH 15](#_Toc112710333)

[2.3. ADOPTION OF CLOUD AND ML IN FINANCIAL SECTOR 16](#_Toc112710334)

[2.4. USES OF CLOUD AND ML IN FINANCIAL SECTOR 18](#_Toc112710335)

[2.4.1. MICROSERVICES 18](#_Toc112710336)

[2.4.2. DIGITAL MARKETING 18](#_Toc112710337)

[2.4.3. CHATBOTS 19](#_Toc112710338)

[2.4.4. DATA COLLECTION 20](#_Toc112710339)

[2.4.5. GDPR 22](#_Toc112710340)

[2.4.6. CUSTOMER SCREENING & CRDIT RISK 25](#_Toc112710341)

[2.4.7. FRAUD DETECTION 26](#_Toc112710342)

[2.5. SAAS & PAAS OFFERINGS 27](#_Toc112710343)

[2.6. SUMMARY 28](#_Toc112710344)

[3. RESEARCH METHODOLOGY 29](#_Toc112710345)

[3.1. INTRODCUTION 29](#_Toc112710346)

[3.2. RESEARCH METHODOLOGY 29](#_Toc112710347)

[3.2.1. DATA SELECTION AND AQUISITION 29](#_Toc112710348)

[3.2.2. DATA RPE-PROCESSING AND TRANSFORMATION 30](#_Toc112710349)

[3.3. PROPOSED METHOD 31](#_Toc112710350)

[3.4. INDEPENDENT VARIABLES TO BE USED FOR REGRESSION 34](#_Toc112710351)

[4. IMPLEMENTATION 37](#_Toc112710352)

[4.1. DATA CLEANSING 37](#_Toc112710353)

[4.1.1. IMPORT HMDA DATA 37](#_Toc112710354)

[4.1.2. REMOVE UNNECESSARY COLUMNS 37](#_Toc112710355)

[4.1.3. CLEAN LOCATION, RACE, ETHNICITY AND CREDIT MODELS 38](#_Toc112710356)

[4.1.4. STANDARDIZING OUTCOMES AND AUTOMATED UNDERWRITING SYSTEM 39](#_Toc112710357)

[4.2. CATEGORIZATION OF DATA 40](#_Toc112710358)

[4.2.1. CATEGORIZING LENDERS 40](#_Toc112710359)

[4.2.2. ADDING METRO DEFINITIONS AND ADDING PROERTY VALUES 41](#_Toc112710360)

[4.2.3. CLEAN-UP AND CATEGORIZE DEBT-TO-INCOME RATIO 41](#_Toc112710361)

[4.2.4. CATEGORIZE AGE AND GENDER OF APPLICANTS 42](#_Toc112710362)

[4.3. REGRESSION 43](#_Toc112710363)

[4.3.1. COLLINEARITY TEST 44](#_Toc112710364)

[4.3.2. CLEAN VARIABLES WITH HIGH VARIANCE INFLOW FACTOR 45](#_Toc112710365)

[4.3.3. REGRESSION 48](#_Toc112710366)

[4.3.4. REGRESSION RESULTS 48](#_Toc112710367)

[4.3.5. ACCURACY OF THE STUDY 52](#_Toc112710368)

[5. OBSERVATIONS AND RECOMMENDATIONS 52](#_Toc112710369)

[5.1. FINDINGS 52](#_Toc112710370)

[5.2. RESULTS OF NYCKEL MODEL TRAINED WITH HMDA DATA 53](#_Toc112710371)

[5.3. RECOMMENDATIONS 54](#_Toc112710372)

[5.4. SUMMARY 55](#_Toc112710373)

[REFERNCES 56](#_Toc112710374)

[6. APPENDIX A: RESEARCH PLAN 61](#_Toc112710375)

[7. APPENDIX B: RESEARCH PROPOSAL 61](#_Toc112710376)

[7.1. BACKGROUND 61](#_Toc112710377)

[7.2. PROBLEM STATEMENT 62](#_Toc112710378)

[7.3. RESEARCH QUESTIONS 63](#_Toc112710379)

[7.4. AIM AND OBJECTIVES 63](#_Toc112710380)

[7.5. SIGNIFICANCE OF THE STUDY 64](#_Toc112710381)

[7.6. SCOPE OF THE STUDY 64](#_Toc112710382)

[7.7. RESEARCH METHODOLOGY 65](#_Toc112710383)

[7.8. RESOURCE REQUIREMENTS 66](#_Toc112710384)

[8. APPENDIX C: ETHICS FORMS 66](#_Toc112710385)

# LIST OF EQUATIONS

[Equation 1: Formula for VIF calculation 45](#_Toc112631119)

# LIST OF TABLES

[Table 1: A table summarizing the results of literature review listing various aspects of FinTech industry and the use of Cloud technologies and AI 28](#_Toc112631079)

[Table 2: List of variables to be considered for applicant segmentation and logistics regression 34](#_Toc112631080)

[Table 3: Description of variables used for regression 34](#_Toc112631081)

[Table 4: List of all independent variables 47](#_Toc112631082)

[Table 5: Regression results 49](#_Toc112631083)

[Table 6: Results of model trained on nyckel.com 55](#_Toc112631084)

[Table 7: Proposed variables for analysis 66](#_Toc112631085)

# LIST OF FIGURES

[Figure 1: PWC poll on AI adoption 8](#_Toc112631131)

[Figure 2: Representation of bias in the data, user interaction and algorithm feedback loop 10](#_Toc112631132)

[Figure 3: Comparison of economies (Source- (Mei, 2022)) 15](#_Toc112631133)

[Figure 4: The evolution of digital marketing (Source- (Riva and Pilotti, 2021; Kamal, 2016)) 19](#_Toc112631134)

[Figure 5: Sample Bot/Digital assistant feature on Royal Bank of Scotland website (Source: Internet) 20](#_Toc112631135)

[Figure 6: Percentage of government websites parenthesis) that contain ≥ 1 cookie per G20 country. (Gotze et al., 2022) 21](#_Toc112631136)

[Figure 7: Percentage of government URLs (number in parenthesis) that contain ≥ 1 cookie per G20 country.(Gotze et al., 2022) 21](#_Toc112631137)

[Figure 8: Comparison of trackers on HSBC bank website in Malta to India (source: www.theMarkup.org) 23](#_Toc112631138)

[Figure 9: Scan of Royal bank of Scotland website in UK and a loan interest rate comparision website in India (source: www.theMarkup.org) 24](#_Toc112631139)

[Figure 10: Schematic diagram of ML in credit risk analysis (Bao, Lianju and Yue, 2019) 26](#_Toc112631140)

[Figure 11: List of services offered by Mambu (Source - company website) 27](#_Toc112631141)

[Figure 12: HMDA Data available on the US government webste (Source: https://ffiec.cfpb.gov/data-publication/dynamic-national-loan-level-dataset/2021) 31](#_Toc112631142)

[Figure 13: Code to convert txt file to a csv file 31](#_Toc112631143)

[Figure 14: Nyckel user interface to upload the training dataset 32](#_Toc112631144)

[Figure 15: Nyckel user interface to define binary logistic regression labels 33](#_Toc112631145)

[Figure 16 Code snippet to clean the location data 40](#_Toc112631146)

[Figure 17: Logic for standardizing outcomes 41](#_Toc112631147)

[Figure 18: Creating white gradient 42](#_Toc112631148)

[Figure 19: Categorization of age 44](#_Toc112631149)

[Figure 20: Removing non-conventional outcomes and income below zero 45](#_Toc112631150)

[Figure 21: List of independent variables 45](#_Toc112631151)

[Figure 22: Calculation of threshold 46](#_Toc112631152)

[Figure 23: Variables with high threshold 46](#_Toc112631153)

[Figure 24: Filtering out high VIF variables and CLTV above 100 47](#_Toc112631154)

[Figure 25: Regression run 49](#_Toc112631155)

[Figure 26: Consolidation of results 52](#_Toc112631156)

[Figure 27: Odds ratios by race 52](#_Toc112631157)

[Figure 28: Odds ratio by debt-to-income ratio 53](#_Toc112631158)

[Figure 29: Confusion matrix 53](#_Toc112631159)

[Figure 30: Graphical representation of odds by race 54](#_Toc112631160)

[Figure 31 : Research plan 62](#_Toc112631161)

LIST OF ABBREVATIONS

AI.............. Artificial Intelligence

API............ Application Programming Interface

CFO........... Chief Finance Officer

CFPB.......... Consumer Financial Protection Bureau

CRM........... Customer Relationship Management

EU............. European Union

FFIEC......... Federal Financial Institutions Examination Council

GDPR......... General Data Protection Regulation

HMDA........ Home Mortgage Disclosure Act

HTTP.......... Hyper Text Transfer Protocol

IAT............ Implicit Association Test

KYC.......... Know your customer

LGBTQ....... Lesbian Gay Bi Trans Queer

ML............ Machine Learning

PAAS......... Planform as a service

SAAS......... Software as a service

US............. United States

USD........... United States Dollar

VIF………. Variance Inflation Factor

CLTV……. Combined Loan to value

# INTRODUCTION

# ABSTRACT

The adoption of Machine Learning and Artificial Intelligence has been on a rise in the past decade. The estimated revenues for Artificial Intelligence services were over 156 billion USD in 2020 (Columbus, 2020). It is also predicted to reach 22 billion USD by 2025, just for financial services industries with an estimate that over half the Financial Services Organizations have already adopted AI for. (Columbus, 2020).

Below is the result of survey conducted by Price-Water-Coopers (PWC, 2021)

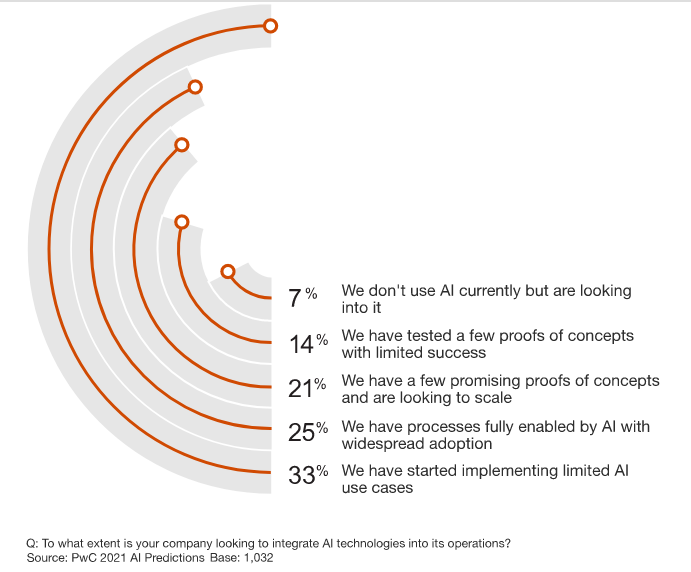


Figure 1: PWC poll on AI adoption

Gartner for Finance recommends Data and analytics among top 10 to-dos for all CFOs(*The Digital Future of Finance Gartner for Finance*, 2021) while also highlighting about the importance of Data Literacy and governance.(*The IT Roadmap for Digital Business Transformation Excerpt*, 2021).

# OVERVIEW OF CLOUD COMPUTING AND MACHINE LEARNING

Though the future looks bright for the applications of AI and ML in most of the industry sector, the biases(Howard and Borenstein, 2018) and the resulting causalities of poorly designed and implemented Algorithms is becoming more common. This is because the regulations and laws on AI are non-existent and most of the implementations are opaque. Even companies well renowned for their technical competence have been unsuccessful in some of their attempts to use Machine learning to take decisions for them. The most infamous case being Amazon’s AMZN.O algorithm for screening resumes showing a bias against women. (Dastin, 2018).

When machine learning is poorly implemented by streaming services to recommend videos or by e-commerce/retail industries to create custom offers for customers, the real-world impact of this is negligible. But when it comes to financial services where small mistakes can lead to Individuals suffering and large mistakes can even lead to economic crashes. With fraud detection and credit screening(Louis Columbus, 2020) being the most used applications of AI, the risk of Machine learning algorithms discriminating against particular groups in the society is also a real possibility.

Research by a non-profit organization called theMarkup on the biases in Mortgage Approval Algorithms used in the United States (Emmanuel Martinez, 2021) has sparked the debate about Algorithmic Accountability. (Martinez & Carollo, 2021).

Good machine learning algorithms need clean data and in ideal cases multiple point of truths. But with regulations over data collection, handing and processing non-existing outside of the European Union’s GDPR law, also with various functions of these financial originations being outsourced or using SAAS and PAAS offerings available in the market, these algorithms are becoming black boxed (Angwin, 2022)

AI systems and algorithms are driven by data on which to be trained. Thus, the functioning of the algorithm is very closely coupled with this data. If the training dataset contains any biases, the algorithms will learn these biases and reflect upon them into their predictions. (Mehrabi *et al.*, 2021). In an ideal world, this should be identified during the test phase after training. But since some of these biases stem from the society itself, it may not be found till it adversely effects a section of the population.

A serious example of this was this was the earlier example of Amazon hiring algorithm learning to favoring Men over Women which is elaborated in the next segment.

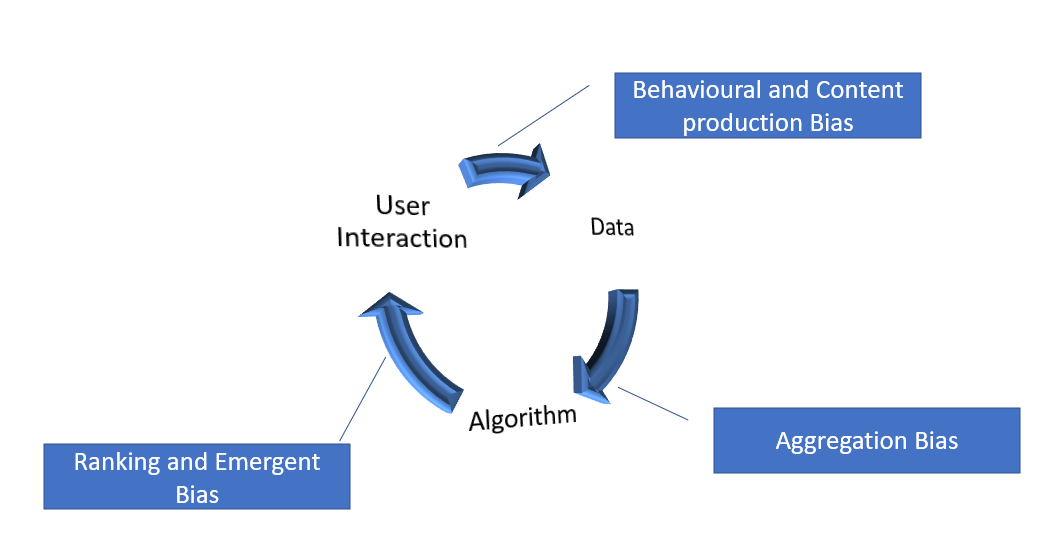


Figure 2: Representation of bias in the data, user interaction and algorithm feedback loop

The real-world biases are often introduced into the data lead to the AI algorithms also exhibiting these biases. (*OpenAI*.org, 2022)

# PROBLEM STATEMENT

There will always be biases in the society and thus the probability of the same biases being introduced into the data is also very high(Ntoutsi *et al.*, 2020; Ferrer *et al.*, 2021). Since, machine learning algorithms are isolated in terms of their perception of the world, they are by design unable to differentiate between correlation and causation.(Mogaji, Soetan and Kieu, 2020) In other words, they do not have common sense. The only world the data that is fed to them. So, they never identify or fill the gaps. (*Bhattacharya, 202*1)

For example, the case of Amazon resume screening algorithm discriminating against female applicants was because the data used to train the algorithm on suitable profiles was based on the resumes of the existing Amazon staff and the lack of diversity that was pre-existing in the organization was seen as women being undesirable for employment by the AI algorithm. (*Dastin, 2018*).

Implicit bias is where a person’s associations and reactions automatically emerge and there is often an unawareness of the presence of said bias(Daumeyer *et al.*, 2019). Project Implicit is a non-profit organization of Harvard researches working since 1998 towards educating public about bias and since 2011 provides a virtual test to assess it online anonymously. It is called Implicit Association Test (IAT) (Tonya R. Moon, 2011) This test is used by many leading organizations such as KPMG include this test as a mandatory training for all new employees. The jury is still out on the effectiveness of such trainings but it certainly does help if not completely eradicate the implicit biases.(Christine Ro, 2021)

Cloud computing and Machine learning are being used for entire customer journey from marketing, lead generation, customer onboarding, services, support and upselling/cross selling(Huttunen *et al.*, 2019).

There is a need to look into how these emerging technologies are being used, existing ethical issues, impact of this on the customer and ways to mitigate these.(Gotthardt *et al.*, 2020)

# AIMS AND OBJECTIVES OF THE RESEARCH

The primary objective of the current research is to study the way technology is being adopted in Fin Serv and propose a method to identify techniques to eliminate biases

The study purposes are formulated as follows:

* To analyze the different areas of business where cloud computing and AI are being used
* To investigate how AI SAAS offerings such as Nyckel and Lobe compare to traditional data science approach
* To suggest ways to identify and predict potential issues in data

# RESEARCH QUESTIONS

The following research questions are suggested for each of the research objective as highlighted as follows.

1. What are the different sub domains in financial services where Cloud computing, AI, ML are used.?
2. What are the different cloud-based offerings available for financial services?
3. How to identity and eliminate biases in Machine learning?

# SCOPE OF THE RESEARCH

The study will comprise the following:

The services offered by the financial services organizations to their customers use Cloud computing, the different data collection techniques used by financial services industries and analyse the Dynamic National Loan-Level Dataset available on FFIEC website to look identify discrimination patterns and find ways to solve them.

The research also aims at ways to identify biases in the datasets and the implications of using such datasets to train and test machine learning models.

# SIGNIFICANCE OF THE RESEARCH

Unlike a decade ago, when technology was mostly limited to people who were literate in computer science, the affordability of smart phones and the internet has led to people from all walks of like becoming exposed to technology. The entry barrier to use the present-day computing machines is very low. Thus, most people who use an app or a website for example do not know how they work, so they cannot be expected to act in their own interests. So, there is a need for research and push for regulation on how technology is used especially in sectors that can make or break people’s lives.

A rejected mortgage or a loan can have a devastating impact on a person and their family. When decisions of such high importance are being delegated to computers, it is important to analyse and measure the accuracy of these decisions.

The current legal system doesn’t hold algorithms accountable for any problems they have caused. For example, when research showed that crime prediction software such as PredPol had consistently discriminated against the minorities in the United States or when tenant screening agencies were found to have wrongly flagged people as having criminal back ground for sharing the same name and surname as someone in the database of convicted felons, there are no real-world consequences to the organizations that sell these products. The false positives on the other hand can have a devastating impact on the individuals at the other end of these automated decision-making algorithms. (Rubel, Castro and Pham, no date; Goodyear, 2022)

The economic impact caused by SARS Covid-19 virus is still reverberating across the world. Many countries helped out its citizens by freezing the mortgage and other loan repayments for a few months in 2020. But now the financial obligations are back to normal even if a number of businesses still suffer. This will have a long-standing impact in the credit history of many more people, especially the underprivileged and poor. The data of last 2 years will affect the decision making of numerous algorithms over the next decade or so. So, the importance of understanding, identifying and resolving the implicit biases of algorithms is even more vital than it has ever been before.

# STRUCTURE OF THE RESEARCH

The study looks into how cloud services and the various phases or departments of financial services industry. It also looks into the Machine learning applications in the industry along with the data collection methods along with the legal and ethical aspects of these methods.

Finally, the plan is to look into the mortgage approval data from the United States government over the last 4 years and look at the any bias or discrimination in the approval process. Further, this data will be fed into cloud-based AI offerings such as Lobe and Nyckel to see if the algorithms show a similar bias in their predictions and compare this to a manual train and testing method using Python with the help of K-Means and Logistic regression methods.

# LITERATURE REVIEW

# OVERVIEW OF FINTECH

Fintech refers to the incorporation of technology into offerings by financial services corporations in order to increase efficiency and delivery to consumers. FinTech mainly works by splitting the products/services of such businesses and creating new markets for them. The number of start-ups and the funding for these start-ups is on the rise but the industry still has a lot of regulatory problems.

The term financial technology applies to any new technical or digital innovation in the way customers and businesses interact with each other. From digital payments to book keeping to background screenings to double-entry bookkeeping. Since the smartphone revolution, the technology in financial sector has grown exponentially and fintech which originally perceived as computerization of back offices of banks and trading companies, now can be describes as a wide variety of technological intervention into commercial and personal finance.

Fintech now covers a vast number of financial activities such as money transfers, credit card or loan applications, ability to update contact information via smartphones, digital wallets, managing investments, all done without the assistance of an employee from the financial institution. According to the 2019 Fintech Adoption Index by the big 4 accounting firm Ernst & Young, over 30% of the customers use at least two or more fintech services and more customers are considering fintech as a part of their daily lives.

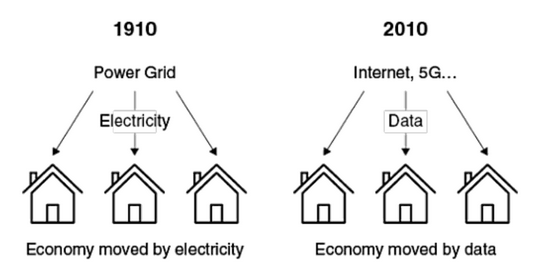


Figure 3: Comparison of economies (Source- (Mei, 2022))

Just like the industrial revolution was powered by electricity, the internet economy is powered by data. Data has much larger significance and impact when it comes to financial services because the real-world impact of the data and decisions made with this data is higher than other areas.

# CLASSIFICATION OF SERVICES AND PRODUCTS IN FINTECH

Below is a non-exhaustive list of services and products that are part of the fintech innovation:

* Cryptocurrency
* Blockchain technology, which is a distributed ledger that maintains records on a network of computers instead of the traditional way of a centralized ledger
* Digital wallets/cash which enable third-party software to integrate financial institutions with bank data. Money management tool Mint is an example of this.
* Smart contracts, which use blockchain to execute contracts between sellers and buyers automatically.
* Self-service banking where mobile applications provide the security to end-users to carry out day-to-day banking operations without every visiting a bank. E.g.: Digital Bank of Singapore allows account opening to operation to closing of the account, all without ever having to talk to or interact with a person.
* Insurance companies using technology to streamline and simplify purchase of policies and processing of claims
* Regulatory technology that helps the financial service firms adhere to government and industry compliance standards. Especially KYC (Know Your Customer) protocols, fraud detection, money-laundering prevention.
* Robo-advisors, using algorithms to automate advice about investments and making it more accessible because its lower in cost.
* Microfinance apps which help under privileged borrow small amounts.
* Cybersecurity. Given the nature of decentralized storage and digital authentication, cyber crime is a very big challenge and thus there are constant innovations in this area.(Qi and Xiao, 2018; Arslanian and Fischer, 2019)

# ADOPTION OF CLOUD AND ML IN FINANCIAL SECTOR

Until about 5 years ago, financial institutions offered their services under a single umbrella, so customers ended up using the same firm for most of their financial needs such as traditional banking activities, trading services, mortgage and investments. In basic terms, Fintech removes the above structure an enables the firms to offer and customers to choose individual products. This enables the fintech companies to cut down costs associated with each transaction by becoming more efficient through streamlining of the offerings.

If one has to define how fintech has affected traditional banking, financial advice, investments and other products, it is ‘disruption’. The branches, salesmen and computers have made way for websites and mobile applications. This, in a way democratizes the financial services from large, monolithic institutions.

For example, Revolut is a mobile and web only banking platform registered in Lithuania which allows users to open a bank account in less than a day. All the verification is done through a mobile app where the users government issues identity card and face are captured using the camera on the smartphone, then verified at the back end to be approved. If you forget the password for the application, the application verifies the user through facial recognition. Any one in the EU special economic zone can open an account and operate it across the EU countries. It also offers international transfers, one-time use virtual debit cards, trading, crypto currency transactions all in one app. To compare it with traditional banking, In Malta getting an appointment to open a bank account with traditional banks such as HSBC and Bank of Valetta can take up to 3 months.

Similarly, stock trading app Robinhood has changed the way equity trading is done. There are peer-to peer lending websites such as Lending Club, business loan providers such as Lendio and Accion. Their popularity and ease of use have all attracted significant investments, such as an Online insurance start-up called Oscar received USD 165 million funding in 2018. This is not an exception; such occurrences have become very common in fintech start-ups.

Thus, we see that more and more financial service institutions are leaning towards adopting fintech as users are also preferring them because of the ease of use and accessibility. (Huang, Gupta and Youn, 2021; Nguyen, Sermpinis and Stasinakis, 2022)

# USES OF CLOUD AND ML IN FINANCIAL SECTOR

# MICROSERVICES

The wide spread use of internet and smartphones has resulted in a drastic change in how users manage their finances. The globalization of banking and investments, the preference to use digital platforms compared to in-person visits to financial institutions added the challenges of security, user experience, reliability and availability to the websites and mobile applications.

Thus, microservices have become the technology of choice to help the organizations to divide and conquer each of these problems while providing a cost-effective way to meet user experience. In fact, use of cloud based micro services is much cheaper than on-premise installation, maintenance of the hardware.

(Nguyen, 2022)

# DIGITAL MARKETING

The emergence of social media and social media marketing with internet companies like Meta and Alphabet that have access to data from billions of users have changed the way lead generation and customer out reach have been managed traditionally.

The modern FinTech companies use may SAAS offerings available to them in the market such as products from Adobe, SalesForce, SAP and other leading CRM solutions to identify and advertise to personalized advertisements.

****

Figure 4: The evolution of digital marketing (Source- (Riva and Pilotti, 2021; Kamal, 2016))

# CHATBOTS

AI powered bots are used both in customer service and also customer support. This redefines the workforce estimations. Lead and sales generation chat bots are equivalent to an employee greeting a customer or a potential who walked into an establishment. They can be designed to enquire and resolve frequently asked queries, collect contact information of a prospect and also direct the user to an expert. All this can be done at a fraction of a cost it would take to employ real persons.

Chatbots also act as a first line of defence in terms of handling volumes to the support team helping the organization manage the customer enquiries and complaints with a relatively less amount of support representatives. But poorly designed chatbots can increase customer frustration and anxiety especially when customers try to contact the financial institutions about time critical events such as identity theft and fraud reporting. (Hoikkala and Ojala, 2022)

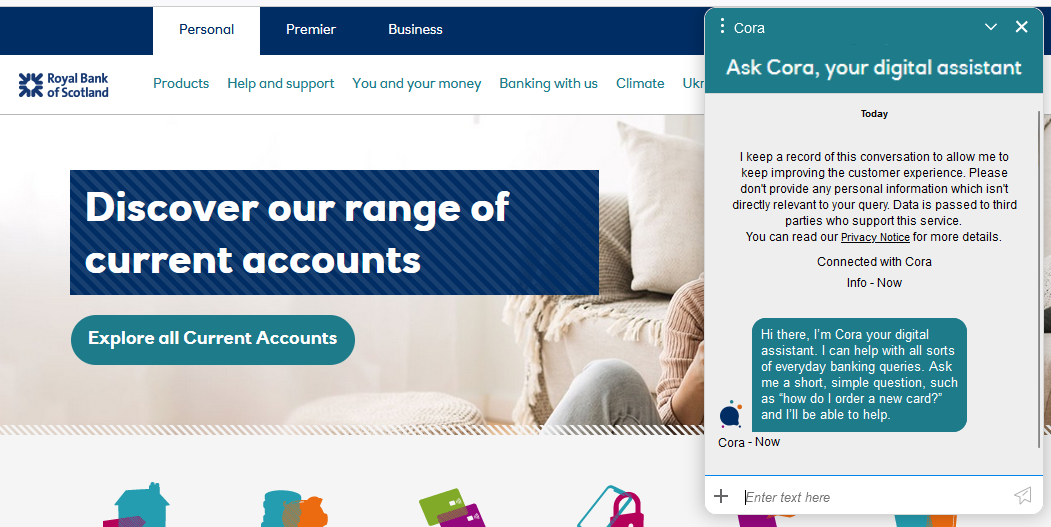


Figure 5: Sample Bot/Digital assistant feature on Royal Bank of Scotland website (Source: Internet)

# DATA COLLECTION

Cookies were originally introduced on websites to enhance the user experience by saving essential information about user so that the stateless HTTP protocol doesn’t force user to identify themselves after every refresh or navigation. The adoption became more common for websites that need login to authenticate and authorize visitors.

The rise of data mining has led to third party cookies which can be embedded onto websites that are used for a lot of purposes. Some are functional like enhancing the performance of webpages, provide analytical data but some are for marketing and tracking.

Toolkits from internet companies such as Alphabet, Microsoft, Adobe Marketing and Meta have incorporated 3rd party cookies websites to collect user data which is then used for marketing purposes.(Freiberg, 2022)

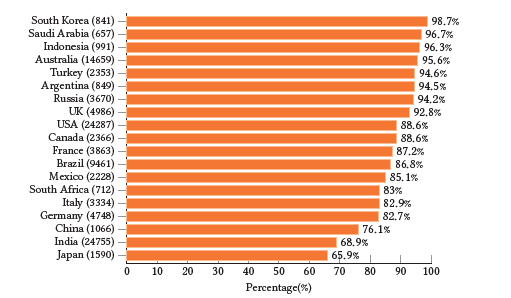


Figure 6: Percentage of government websites parenthesis) that contain ≥ 1 cookie per G20 country. (Gotze et al., 2022)

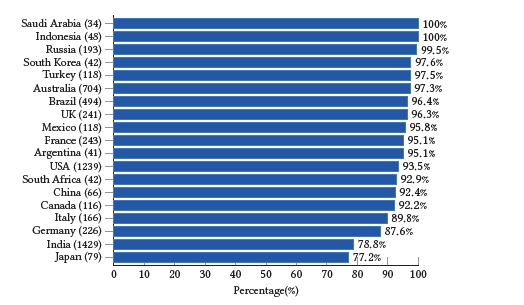


Figure 7: Percentage of government URLs (number in parenthesis) that contain ≥ 1 cookie per G20 country.(Gotze et al., 2022)

Some browsers such as Firefox give the end users the ability to block third party cookies but because the use of internet is not limited to computer literate users and because of the use of browsers such as google chrome and smart phone operating system Android is far greater than any other alternatives available. It lets big tech firms such as Meta and Alphabet to track and identify individuals using these cookies on websites and through tracking mechanisms in mobile applications. This data is then purchased and processed by data brokers who can correlate information from various such sources to identify individuals very specifically with the help of advertising IDs on mobile phones. Facebook estimates a loss of USD 10 billion due to Apple changing the cross-app tracking to opt in from opt out on IOS powered devices.

This may put vulnerable sections of society such as LGBTQ and women seeking abortions (with the criminalization of abortion in the United States) at the risk of being targeted but agenda driven groups. (Madhusudhan and Surashe, 2022).

# GDPR

There is some respite in form of legislations such as GDPR which set strict rules and guidelines for the collection and processing of data. These laws limit the ability of websites to legally collect data without user consent.(Arango Kure, 2022)

Below are the results of a scan by a tool called BlackLight developed by a non-profit organization called TheMarkup. A comparison of the same bank website in Malta to India shows the differences in the trackers and cookies. This is a result of GDPR legislation.

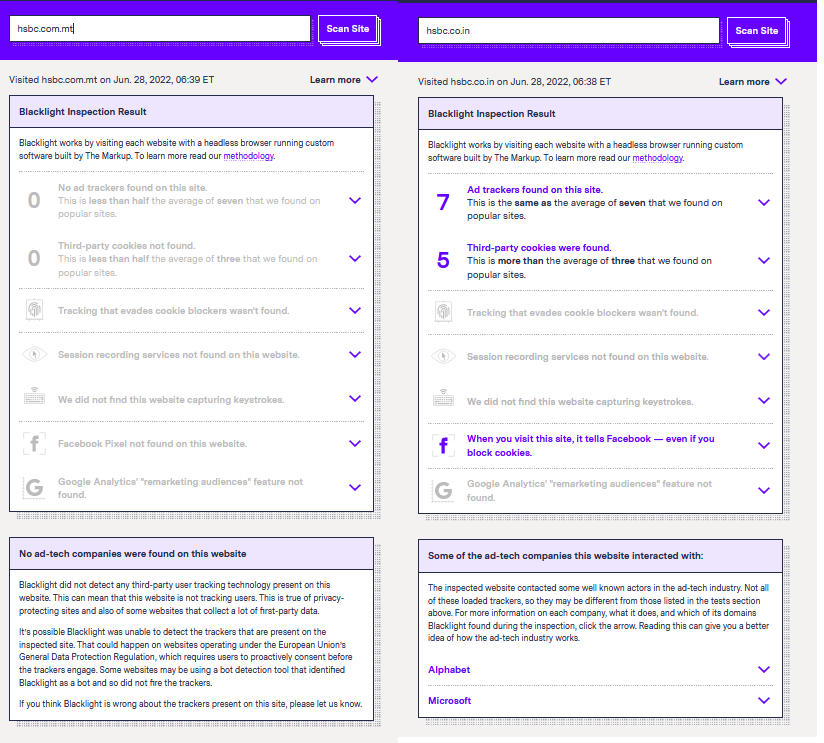


Figure 8: Comparison of trackers on HSBC bank website in Malta to India (source: www.theMarkup.org)

Since, the user content is vital to legally collect data. Some organizations resort to using dark patterns in their user interfaces. These include making it difficult to choose the non-preferred option such as declining consent to tricking users into clicking the wrong button with instructions that are designed to confuse the user. (Waldman, 2020; Luguri and Strahilevitz, 2021)

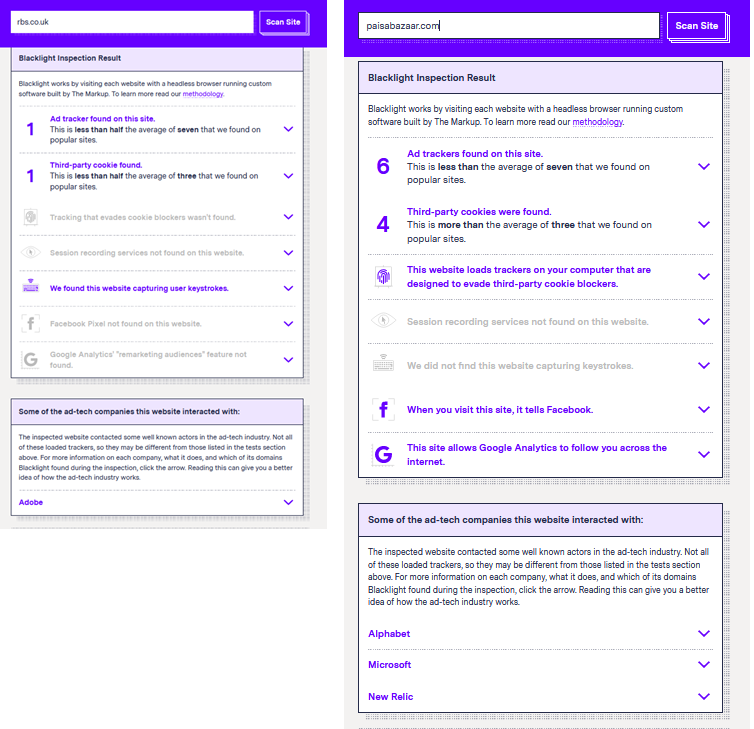


Figure 9: Scan of Royal bank of Scotland website in UK and a loan interest rate comparision website in India (source: www.theMarkup.org)

There are even websites that can capture and send information of key strokes of users without having to

# CUSTOMER SCREENING & CRDIT RISK

The organizations in the business of lending money use various supervised and unsupervised machine learning algorithms to predict the risk of a credit line with the applicant.

In traditional risk modelling, an existing customer is trusted more than a new customer. This limits the growth of business as companies adopted a safety-first approach. Even in case of existing customers, the customer segmentation took time because the banks couldn’t trust the customer until they had a reasonably long relationship. Such strategies were not fool proof as there were always cases of defaulters. But the data gathering and mining by data brokers and their partnerships with entities such as TransUnion CIBIL has enabled the financial institutions to outsource the risk analysis. The collaboration of all financial institutions to share data of their customers and their credit history with them helps other institutions make informed decisions when evaluating applicants.

Another approach is to use deep learning and AI to train an algorithm with existing credit history along with other parameters that can influence a decision and let the algorithm make decisions. (Bao, Lianju and Yue, 2019)

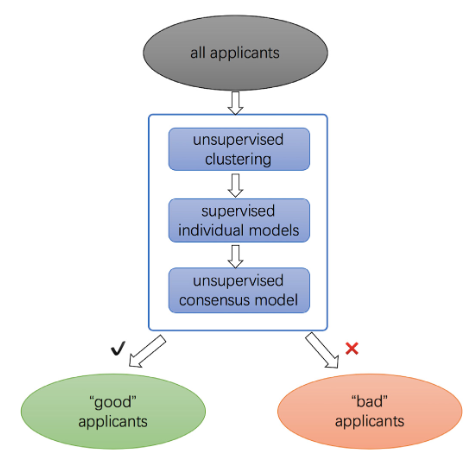


Figure 10: Schematic diagram of ML in credit risk analysis (Bao, Lianju and Yue, 2019)

The main focus of this research is to look at the different supervised and unsupervised models and see the different implicit biases that can be introduced into these algorithms, their real-world impact and ways to mitigate these issues.

# FRAUD DETECTION

Credit card companies and banks use AI algorithms to monitor transactions to flag or block suspicious transactions. Such algorithms generally use a combination of customer data with a database of historical fraud reports to identify any transactions that are deemed suspicious. For example: A customer in living USA initiating a transaction from Europe is a potential fraudulent transaction. Though there are always cases where genuine transactions may be flagged, safety first approach is also preferred by customers, so these algorithms do not have a problem with false positives.(Awoyemi, Adetunmbi and Oluwadare, 2017)

# SAAS & PAAS OFFERINGS

The SAAS and PAAS offerings in the domain of cloud computing and ML have also contributed to the higher rates of adoption of cloud and AL/ML. A lot of start-ups rely on such products instead of developing and maintaining these systems inhouse.

The tried and tested offerings thus lower the entry barrier required to use these technologies.

Even in case of companies wanting to develop their own algorithms, there are some services such as Lobe and Nyckel which provide a APIs to machine learning models on a subscription basis.

Some platforms such as Mambu go further and provide the entire eco-system of products needed for a digital banking company on subscription. (Pantielieieva *et al.*, 2018)

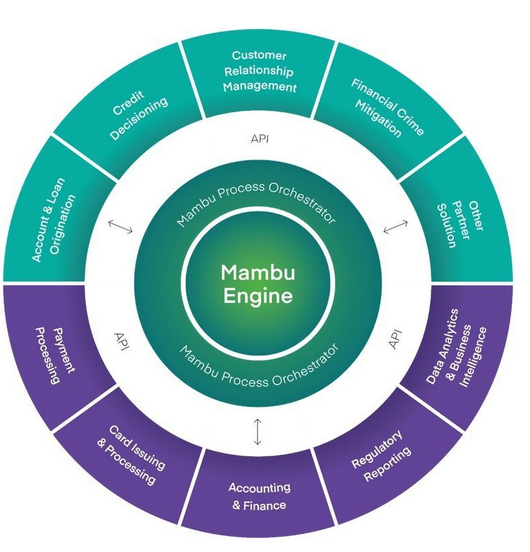


Figure 11: List of services offered by Mambu (Source - company website)

# SUMMARY

Table 1: A table summarizing the results of literature review listing various aspects of FinTech industry and the use of Cloud technologies and AI

|  |  |  |
| --- | --- | --- |
|  | **Cloud** | **Machine learning & AI** |
| **Data collection** | Yes | Not applicable |
| **Data storage** | Yes | Not applicable |
| **Customer support** | Yes | Yes |
| **Sales and lead generation** | Yes | Yes |
| **Risk analysis** | Yes | Yes |
| **Background checks** | Yes | Yes |
| **Digital platforms** | Yes |  |
| **Security** | Yes | Yes |
| **Fraud detection** | Yes | Yes |

# RESEARCH METHODOLOGY

# INTRODCUTION

# RESEARCH METHODOLOGY

# DATA SELECTION AND AQUISITION

The United States Home Mortgage Disclosure Act of 1975 enables the lending data across America. Traditionally, this data has been relied upon to identify any lending disparities. It was done to ensure that the needs of communities are being met by the lenders that serve them and to identify any discriminatory practices. This is currently maintained by CFPB.

Over the years, Federal government has mandated that the lenders report data on mortgage applications and prospective customers. In response to the housing market collapse in 2008, Dodd-Frank Act was passed in 2010 to expand this. A series of reforms were enacted through this federal legislation, among them was a rule to disclose more data, especially the combined loan to value and debt to income ratios. It was 2020 by the time the disclosure rules became effective and were first applied to mortgage data from 2018. This public database has the names and addresses of the applicants redacted. Though the lenders are also required to include the credit scores of the applicants, the government doesn’t release them citing privacy issues.

This HMDA data is loan data that contains details of many individual mortgage applications. The 2020 data set, that was last updated in August 2021 contains more than 1.70 crore applications which is about 90% of all loans made in the United States and from more than 5500 lending institutions. Micro or Small loan lenders are not required to report their data.

The dataset contains detailed information about the loan, lender and property. It contains the bowers details such as sex, income and race among other characteristics. It also details if the mortgage is insured by the government, the means of application, reason for which the mortgage is sought, amount of loan Property and its location.

The dataset contains more than 50 columns, which makes the 2018 & 2019 datasets most expansive ever. This research aims to make use of both these datasets. Dodd-Frank Act has increased the number of fields available in public dataset, most significant of them being debt to income ratio which according to regulators should be the most important factor in considering lending. Only 10 % said that they were concerned about credit scores. (https://www.fico.com/en/newsroom/housing-bubble-inflating-mortgage-lenders-tell-fico, 2014). The CFPB 2019 report arrived at the same conclusion when the looked at why applicants were denied. The agency reported that debt to income ratio was the most common factor for rejection of applications.(*The Markup*, 2021)

This ratio is calculated by dividing the total monthly payments on all open lines of credit by their monthly income. A high ratio means that the applicant uses most of their income in paying debts.(Consumer Response Annual Report, 2020)

# DATA RPE-PROCESSING AND TRANSFORMATION

Government researchers, regulators and analysts often treat conventional loans separately to the mortgages related to Federal housing administration and other loans that are backed by the government (E.g.: US Veteran affairs). Conventional mortgages give a better picture of practices of mortgage industry as these decisions are without government influence.

Limiting the analysis of these specific loans and removing data from government backed loans reduces the data to 24 lakhs in 2019 and 28 lakhs in 2018.

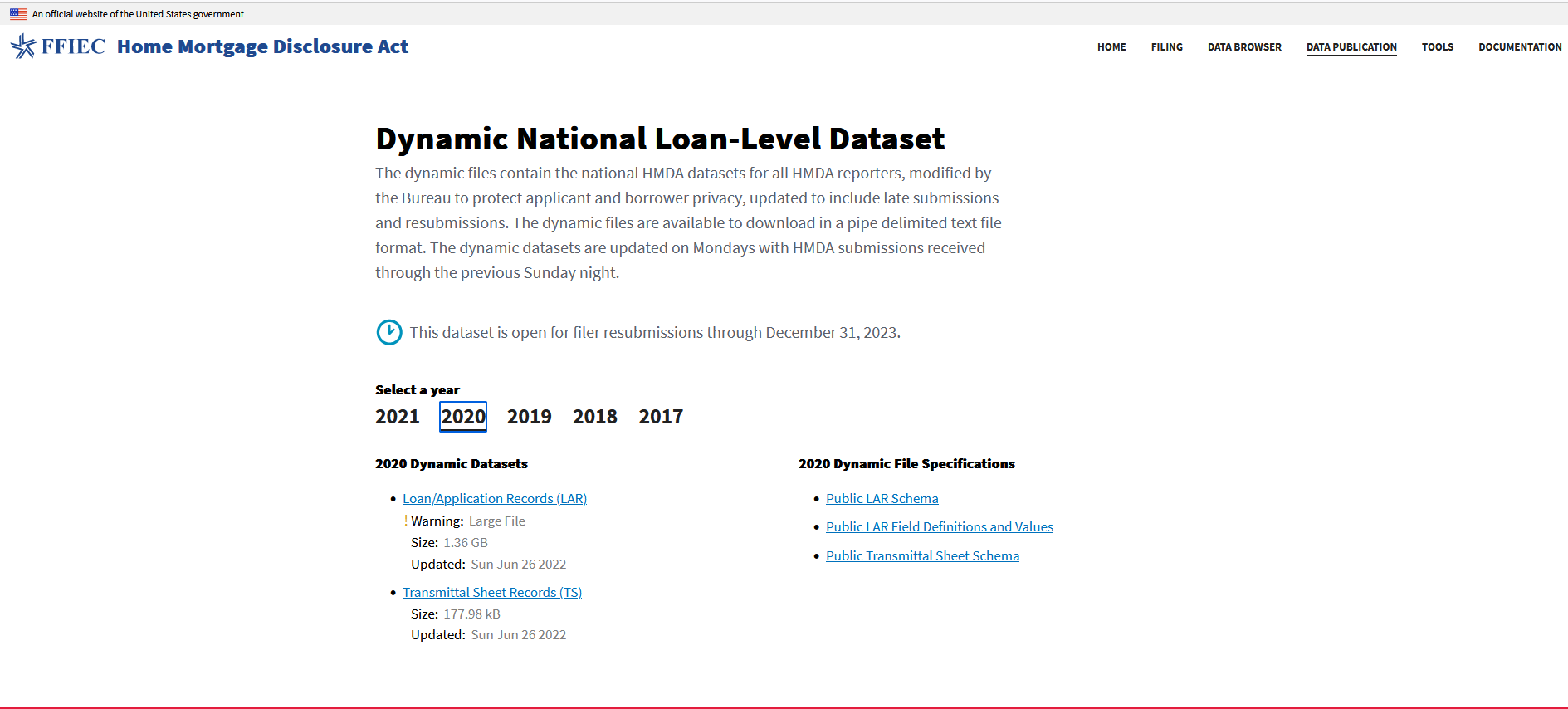


Figure 12: HMDA Data available on the US government webste (Source: <https://ffiec.cfpb.gov/data-publication/dynamic-national-loan-level-dataset/2021>)

The data is available as a zipped text file with forward slash (/) used as a delimiter. The file must be first converted to a format that is readable by a program.

Text

Description automatically generated

Figure 13: Code to convert txt file to a csv file

# PROPOSED METHOD

To determine the general customer segmentation of the data that would generally classify the customers into buckets, the proposed method is to use K-Means algorithm. Then to verify if the debt-income ratio explains the disparities between loan approval rates between different races and ethnicities, the proposed method is binary logistics regression. This type of regression enables the establish relationship between different variables against an outcome. In this case, the outcome is weather the loan was approved or not.

The research also aims at looking at an AI/ML product for machine learning called Nyckel which advertises that the users can build AI into their products without a ML team and having maintain infrastructure.

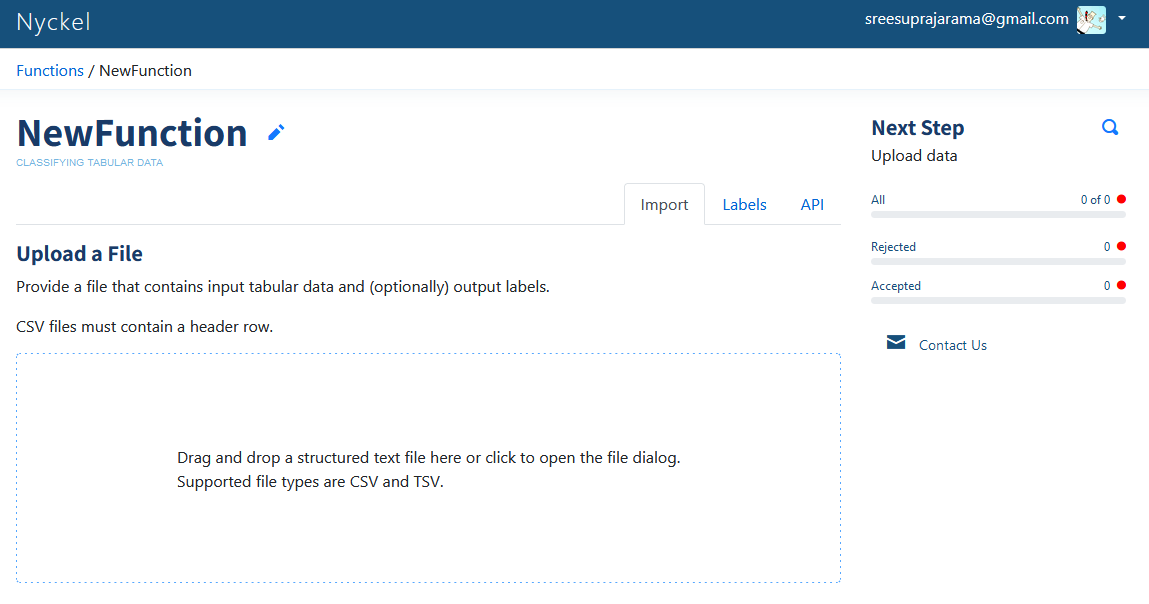


Figure 14: Nyckel user interface to upload the training dataset

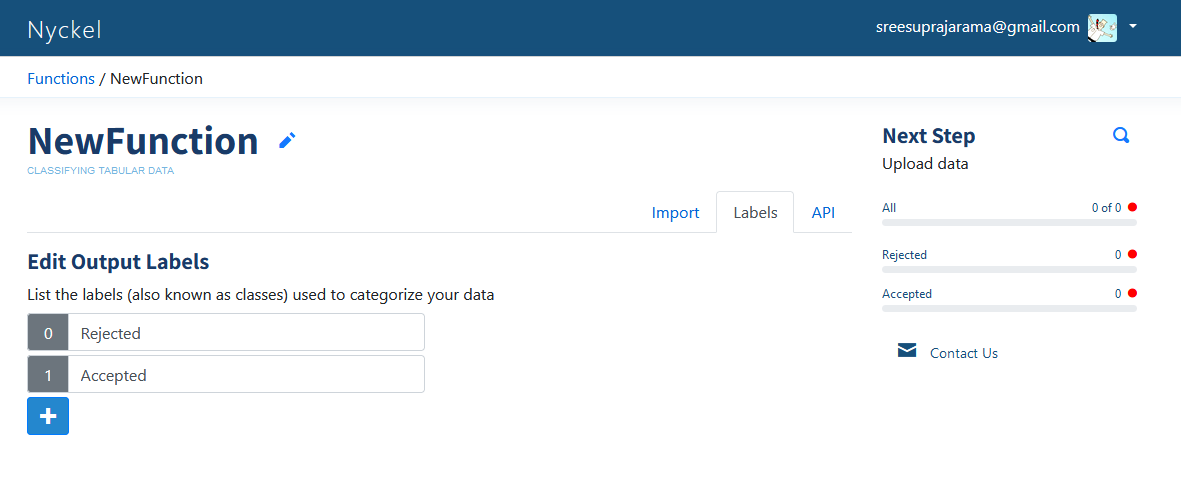


Figure 15: Nyckel user interface to define binary logistic regression labels

The variables to be considered are as follows

|  |  |
| --- | --- |
| 1 | Age |
| 2 | Combined loan-to-value ratio |
| 3 | Credit model used |
| 4 | Debt-to-income ratio |
| 5 | Income |
| 6 | Loan amount |
| 7 | Mortgage term |
| 8 | Non-Hispanic White population percentage of the census tract where the property is located (added variable to HMDA data) |
| 9 | Property value |
| 10 | Race |
| 11 | Sex |
| 12 | Size of the metro area where the property is located (added variable to HMDA data) |
| 13 | The automated underwriting system used |
| 14 | The ratio between the median income of the census tract where the property is located and the median income of the metro area |
| 15 | The size of the lender (added variable to HMDA data) |
| 16 | The type of lender (added variable to HMDA data) |
| 17 | Whether the application had a co-applicant |

Table 2: List of variables to be considered for applicant segmentation and logistics regression

# INDEPENDENT VARIABLES TO BE USED FOR REGRESSION

The independent variables in regression model.

Table 3: Description of variables used for regression

|  |  |  |
| --- | --- | --- |
| **VARIABLE** | **DESCRIPTION** | **VALUES** |
| Race and Ethnicity | The data contains 5 columns related to race this is to accommodate the scenarios where the applicant and co-applicant are from different races.  The data where race information is not available is grouped seperately | |  | | --- | | * Asian/Pacific Islander | | * Black | | * Latino | | * Native American | | * Race not applicable | | * White | |
| Sex | In addition to male and female a separate category was used for data that was missing gender information or had both male and female checked as yes. | * Male * Female * Not applicable |
| Co-applicant | The dataset doesn’t contain a dedicated columns to identify if a loan had a co-applicant. It needs to be derived from the other co-applicant information available such as ethnicity. | * Co-applicant * No co-applicant * Not applicable |
| Age | The data doesn’t contain specific age of applicants but instead classifies applicants into age brackets in the increments of 10. The data contains the classification of 65 to 74 and 75 or older but due to very low number of applicants that fell in the category of 75 or older, last two buckets have been combined to 65 or older. | * < 25 * >= 25 and <34 * >= 35 and <44 * >= 45 and <54 * >= 55 and <64 * >65 * Age not provided |
| Income | Income represents the annual income of the applicant. The data was skewed towards lower end of income with some outliers towards higher side. |  |
| Loan Amount | The loan amount was also considered because of the high amount of gap between incomes of applicants. |  |
| Property Value Ratio | Property value ratio represents requested loan amount to property total value. |  |
| Mortgage term | The data represents Mortgage in terms of number of years as a integer. This will be classified into 4 groups with Not Applicable representing the data that was missing the Mortgage term. | * =30-yrs. * < 30 yrs. * > 30 yrs. * Not applicable |
| Credit Model | The credit score information is missing from the data but the credit model used to calculate credit score is available. There are 10 different models including where credit score was exempt for or missing entirely. |  |
| Debt-to-Income Ratio | The Debt-to-income ratio together with the loan amount are the only data points that need to have an impact on the outcome of the loan application. Based on the percentage available in the data, this will be sorted into 4 different buckets for regression. | * Healthy: a debt-to-income ratio of 35 percent or less * Manageable: a ratio between 36 and 42 percent * Nearing Manageable: a ratio between 43 and 49 percent * Struggling: 50 percent or more * DTI Not Applicable |

# IMPLEMENTATION

The code and python notebooks used for the research can be found at <https://github.com/sumitranandan-k/UpgradThesis>

# DATA CLEANSING

# IMPORT HMDA DATA

The data is available with forward slash as the delimiter and as a text file. The file was converted to a csv file with comma as a delimiter using python code. This file was imported using the Pandas libraries.

* Data can be download from the CFPB site: [2021 LAR dataset](https://ffiec.cfpb.gov/data-publication/dynamic-national-loan-level-dataset/2021)
* The date the file was downloaded was appended to the raw file name.
* 99 Columns
* 26,204,358 records
* [Data Dictionary](https://ffiec.cfpb.gov/documentation/2021/lar-data-fields/)

# REMOVE UNNECESSARY COLUMNS

There are 21 columns that can be dropped as they are either not relevant to the scope of the research or redundant.

The following columns were added by the CFPB, not using them.

* derived\_loan\_product\_type
* derived\_dwelling\_category
* derived\_ethnicity
* derived\_race
* derived\_sex

Focusing on the applicant's first ethnicity

* applicant\_ethnicity-2
* applicant\_ethnicity-3
* applicant\_ethnicity-4
* applicant\_ethnicity-5

Focusing on the co-applicant's first ethnicity. Don't need these columns to find co-applicants.

* co-applicant\_ethnicity-2
* co-applicant\_ethnicity-3
* co-applicant\_ethnicity-4
* co-applicant\_ethnicity-5

Focusing on the applicant's first race

* applicant\_race-2
* applicant\_race-3
* applicant\_race-4
* applicant\_race-5

Focusing on the co-applicant's first race. Don't need these columns to find co-applicants.

* co-applicant\_race-2
* co-applicant\_race-3
* co-applicant\_race-4
* co-applicant\_race-5

# CLEAN LOCATION, RACE, ETHNICITY AND CREDIT MODELS

We first replace all the Nulls with a specific text value to keep the distinction with the data where location is not applicable and group all the country codes and census tract. A similar approach is followed for rest of the independent variables where the values are grouped and transformed before the data with Nulls is dropped and the original columns is deleted once it is processed.

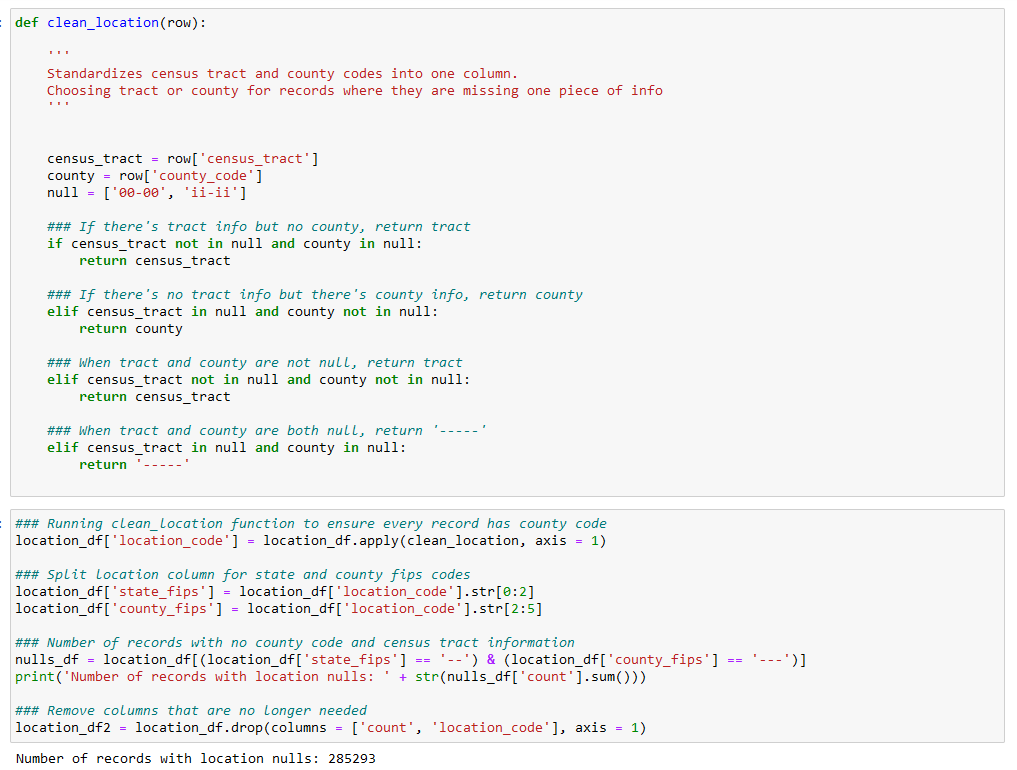


Figure 16 Code snippet to clean the location data

# STANDARDIZING OUTCOMES AND AUTOMATED UNDERWRITING SYSTEM

The data has 8 different outcomes, to simplify the research, they are reduced into 4 buckets.

1: Loans

3: Denials

4: Other Outcomes

6: Purchase loans



Figure 17: Logic for standardizing outcomes

The data of purchase loans is dropped as it is not in scope for this research.

# CATEGORIZATION OF DATA

The cleaning of data has reduced the data set. The number of rows left is not 17,545,457 with 86 columns.

# CATEGORIZING LENDERS

Since the data contains lender codes, we first join it with lender data available on the HMDA website to help categorize them. The data with lender information is dropped and the left are categorized into

1: Banks

2: Credit Union

3: Independent Mortgage Companies

4: No definition

# ADDING METRO DEFINITIONS AND ADDING PROERTY VALUES

Joined the data with standard county codes information to classify the location of property into Metro and non-Metro. Based on the county codes, this data is then joined with the property values data to derive the median value of property based on the GEO\_ID on the loan data. Using census data, the demographic per location per race and ethnicity is added to the data.

A field is not added to define a gradient value for the white population of the county and then merged with the original data frame.

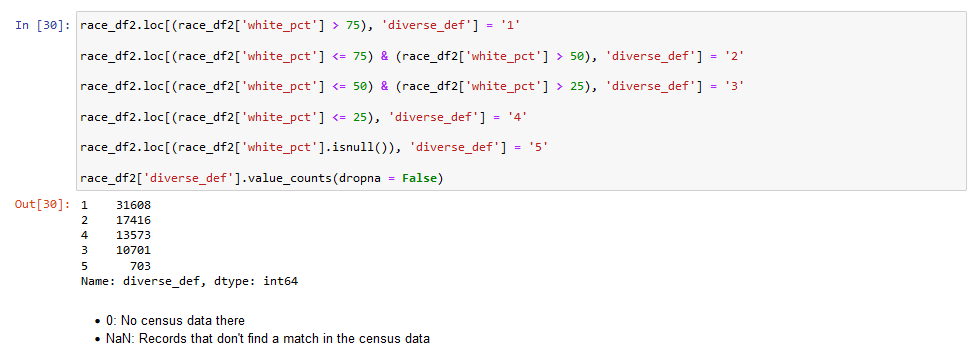
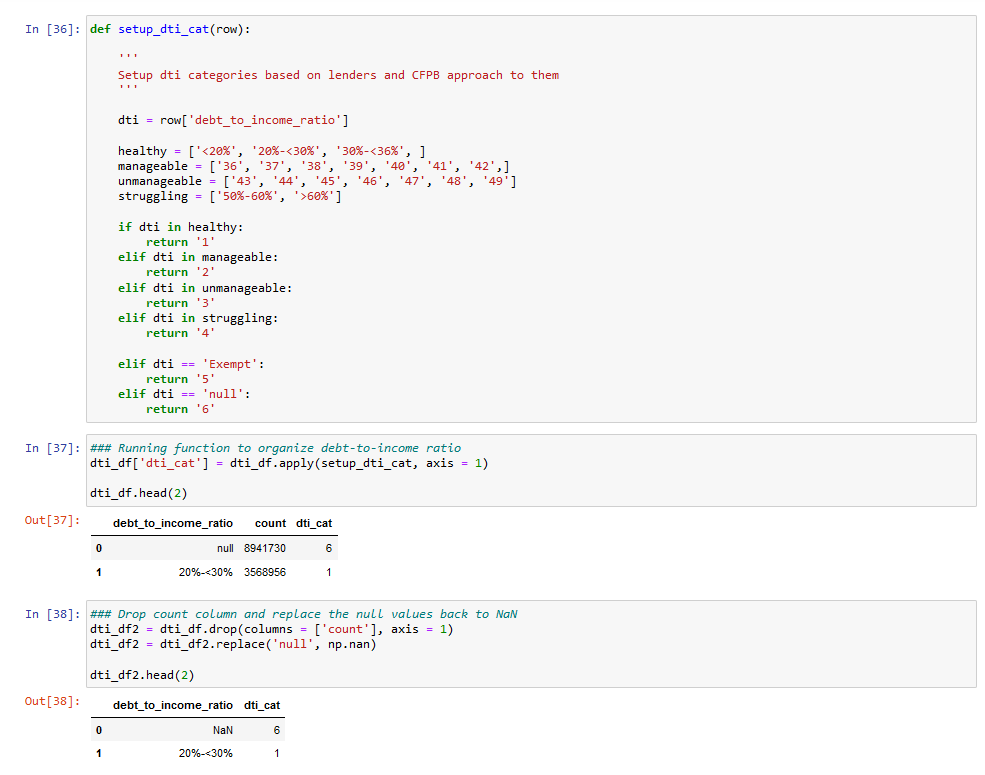


Figure 18: Creating white gradient

# CLEAN-UP AND CATEGORIZE DEBT-TO-INCOME RATIO

The debt-to-income ratio is categorized into healthy, manageable, unmanageable and struggling based on the percentages of the value in the data and all the records with no income available or DTI available are removed from the data frame.



Similar exercise is done with Loan to value and Property values.

# CATEGORIZE AGE AND GENDER OF APPLICANTS

The ages in the data are an integer value. For use in regression, this data is classified into buckets with the size of 10 years.

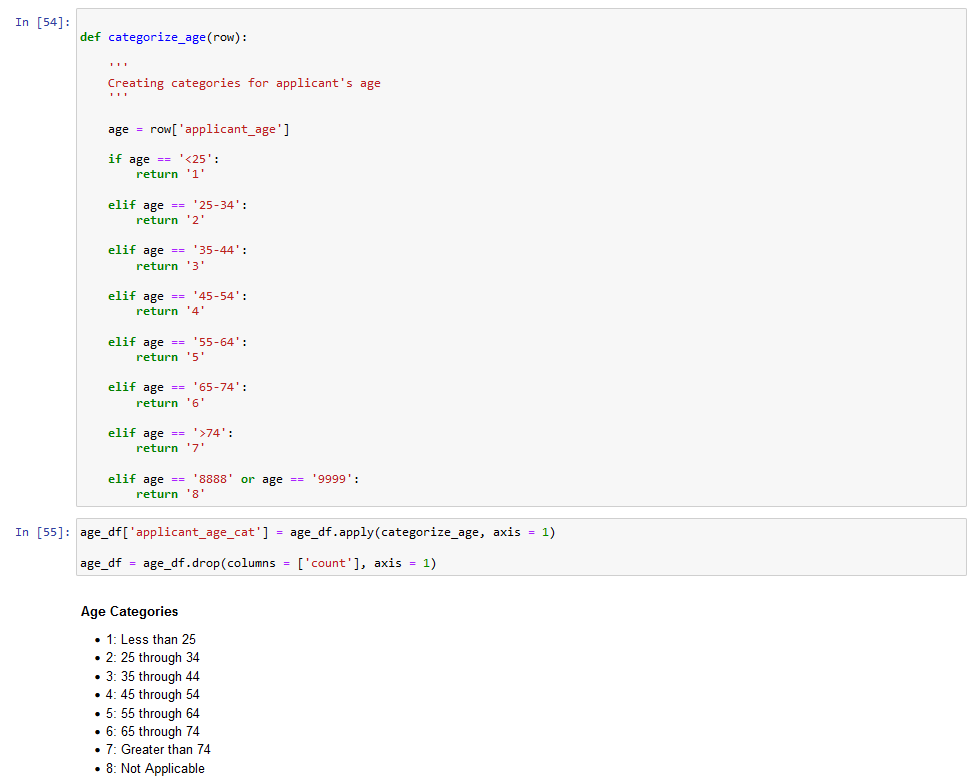


Figure 19: Categorization of age

Similarly, the gender is classified into male and female. All the data that is missing gender information is marked as Not Applicable.

# REGRESSION

The categorized data after cleansing is now 5,329,538 with 117 columns. For the sake of academic research, only the outcomes of approval and rejection are considered. The data with other outcomes is filtered out along with data with income less than or equal to zero. This is to ensure that the Debt-To-Income ratios value is finite.

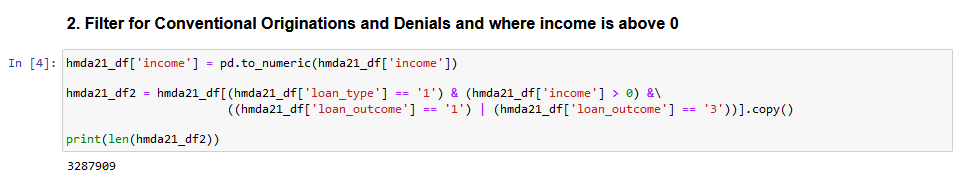


Figure 20: Removing non-conventional outcomes and income below zero

The next step is to define the regression variables and continuous variables. All the variables that were categorized in the Categorization step are defined as regression variables. The regression variables are outcome, ethnicity of applicant and co-applicant, age category, property value, mortgage term, credit model used, debt to income ratio category, down payment percentage category, lender, metro percentile and white gradient. Variables income, loan adequacy requirement value, property value are defined as continuous variables as they were not categorized into buckets.

# COLLINEARITY TEST

The columns that were derived from the exiting data in clean-up steps are defined as independent variables for the regression.



Figure 21: List of independent variables

Some of these variables may not be suitable for regression as their VIF may be above the tolerance limit. The formula used for VIF calculation is inverse of tolerance.

Equation 1: Formula for VIF calculation



Figure 22: Calculation of threshold

The threshold is set as 2.5.

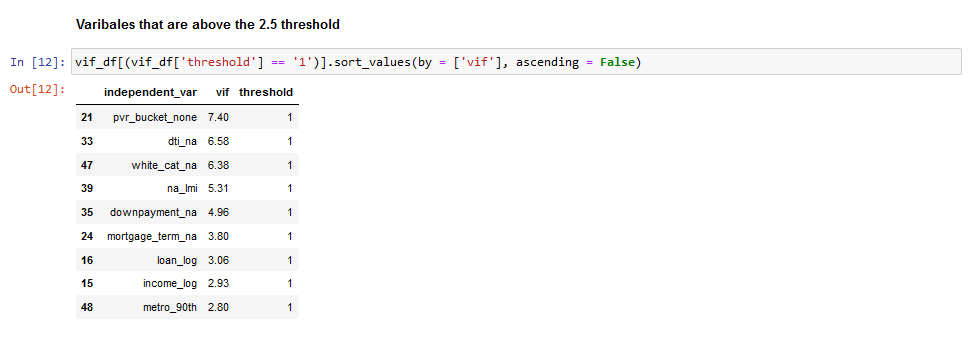


Figure 23: Variables with high threshold

# CLEAN VARIABLES WITH HIGH VARIANCE INFLOW FACTOR

In spite of the high VIF values, income and metro percentile are kept because they are very relevant for the regression. Rest of the independent variables are removed from the list.

The data entries with CLTV greater than 100 are also removed because these indicate the applications that are for a loan value that is greater than the property value. These have high chance of rejection and would skew the regression if included. This exercise leaves us with 51 independent variables.



Figure 24: Filtering out high VIF variables and CLTV above 100

Table 4: List of all independent variables

|  |  |
| --- | --- |
| Serial number | independent variable |
| 1 | black |
| 2 | latino |
| 3 | asian\_cb |
| 4 | native |
| 5 | race\_na |
| 6 | no\_coapplicant |
| 7 | na\_coapplicant |
| 8 | female |
| 9 | sex\_na |
| 10 | less\_than25 |
| 11 | between25\_34 |
| 12 | between45\_54 |
| 13 | between55\_64 |
| 14 | older\_than65 |
| 15 | age\_na |
| 16 | income\_log |
| 17 | loan\_log |
| 18 | property\_value\_ratio |
| 19 | less30yrs\_mortgage |
| 20 | more30yrs\_mortgage |
| 21 | equifax |
| 22 | experian |
| 23 | other\_model |
| 24 | more\_than\_one |
| 25 | model\_na |
| 26 | dti\_manageable |
| 27 | dti\_unmanageable |
| 28 | dti\_struggling |
| 29 | combined\_loan\_to\_value\_ratio |
| 30 | moderate\_lmi |
| 31 | middle\_lmi |
| 32 | low\_lmi |
| 33 | credit\_union |
| 34 | independent |
| 35 | lender\_na |
| 36 | lar\_count |
| 37 | white\_cat2 |
| 38 | white\_cat3 |
| 39 | white\_cat4 |
| 40 | metro\_90th |
| 41 | metro\_80th |
| 42 | metro\_70th |
| 43 | metro\_60th |
| 44 | metro\_50th |
| 45 | metro\_40th |
| 46 | metro\_30th |
| 47 | metro\_20th |
| 48 | metro\_10th |
| 49 | metro\_less10th |
| 50 | micro\_area |
| 51 | metro\_none |

# REGRESSION

Logistic regression was run using the statsmodels libraries.

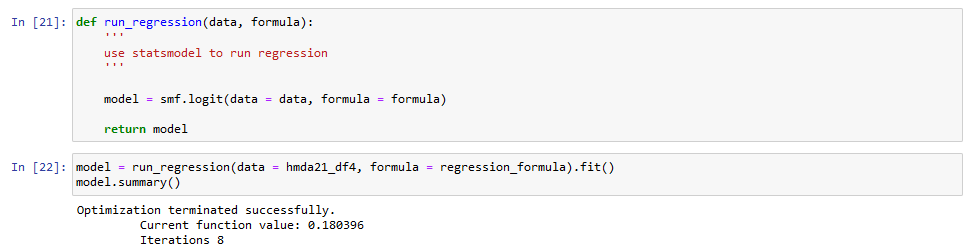


Figure 25: Regression run

# REGRESSION RESULTS

Table 5: Regression results

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dep. Variable:** | denied | | | **No. Observations:** | | | 3112469 | | |
| **Model:** | Logit | | | **Df Residuals:** | | | 3112417 | | |
| **Method:** | MLE | | | **Df Model:** | | | 51 | | |
| **Date:** | Sat, 27 Aug 2022 | | | **Pseudo R-squ.:** | | | 0.219 | | |
| **Time:** | 21:19:36 | | | **Log-Likelihood:** | | | -561480.00 | | |
| **converged:** | TRUE | | | **LL-Null:** | | | -718940.00 | | |
| **Covariance Type:** | nonrobust | | | **LLR p-value:** | | | 0 | | |
| **Variable** | | **coef** | **std err** | | **z** | **P>|z|** | | **[0.025** | **0.975]** |
| Intercept | | -10.10 | 0.30 | | -33.52 | 0.00 | | -10.69 | -9.51 |
| black | | 0.67 | 0.01 | | 63.53 | 0.00 | | 0.65 | 0.69 |
| latino | | 0.42 | 0.01 | | 47.74 | 0.00 | | 0.40 | 0.44 |
| asian\_cb | | 0.41 | 0.01 | | 42.26 | 0.00 | | 0.39 | 0.43 |
| native | | 0.52 | 0.04 | | 14.95 | 0.00 | | 0.45 | 0.59 |
| race\_na | | 0.42 | 0.01 | | 40.91 | 0.00 | | 0.40 | 0.44 |
| no\_coapplicant | | 0.26 | 0.01 | | 42.13 | 0.00 | | 0.25 | 0.28 |
| na\_coapplicant | | 0.40 | 0.11 | | 3.58 | 0.00 | | 0.18 | 0.61 |
| female | | -0.08 | 0.01 | | -13.34 | 0.00 | | -0.09 | -0.07 |
| sex\_na | | 0.20 | 0.01 | | 15.41 | 0.00 | | 0.18 | 0.23 |
| less\_than25 | | -0.03 | 0.01 | | -2.18 | 0.03 | | -0.06 | 0.00 |
| between25\_34 | | -0.21 | 0.01 | | -28.31 | 0.00 | | -0.22 | -0.19 |
| between45\_54 | | 0.14 | 0.01 | | 17.50 | 0.00 | | 0.13 | 0.16 |
| between55\_64 | | 0.11 | 0.01 | | 11.55 | 0.00 | | 0.09 | 0.13 |
| older\_than65 | | -0.08 | 0.01 | | -6.46 | 0.00 | | -0.10 | -0.05 |
| age\_na | | 1.68 | 0.24 | | 7.15 | 0.00 | | 1.22 | 2.14 |
| income\_log | | -0.12 | 0.01 | | -20.01 | 0.00 | | -0.14 | -0.11 |
| loan\_log | | 0.47 | 0.01 | | 69.66 | 0.00 | | 0.46 | 0.48 |
| property\_value\_ratio | | -0.06 | 0.00 | | -29.11 | 0.00 | | -0.07 | -0.06 |
| less30yrs\_mortgage | | -0.02 | 0.01 | | -1.56 | 0.12 | | -0.04 | 0.00 |
| more30yrs\_mortgage | | 0.46 | 0.02 | | 20.27 | 0.00 | | 0.42 | 0.50 |
| equifax | | -0.04 | 0.01 | | -5.16 | 0.00 | | -0.05 | -0.02 |
| experian | | 0.05 | 0.01 | | 6.48 | 0.00 | | 0.03 | 0.06 |
| other\_model | | 0.86 | 0.02 | | 44.37 | 0.00 | | 0.83 | 0.90 |
| more\_than\_one | | -0.25 | 0.02 | | -13.78 | 0.00 | | -0.28 | -0.21 |
| model\_na | | 0.28 | 0.01 | | 25.15 | 0.00 | | 0.26 | 0.30 |
| dti\_manageable | | 0.01 | 0.01 | | 1.59 | 0.11 | | 0.00 | 0.03 |
| dti\_unmanageable | | 0.30 | 0.01 | | 39.95 | 0.00 | | 0.29 | 0.32 |
| dti\_struggling | | 4.31 | 0.01 | | 406.73 | 0.00 | | 4.28 | 4.33 |
| combined\_loan\_to\_value\_ratio | | -0.01 | 0.00 | | -50.20 | 0.00 | | -0.01 | -0.01 |
| moderate\_lmi | | 0.08 | 0.01 | | 8.85 | 0.00 | | 0.06 | 0.10 |
| middle\_lmi | | 0.02 | 0.01 | | 2.79 | 0.01 | | 0.01 | 0.03 |
| low\_lmi | | 0.15 | 0.02 | | 8.89 | 0.00 | | 0.12 | 0.19 |
| credit\_union | | 0.04 | 0.01 | | 4.02 | 0.00 | | 0.02 | 0.06 |
| independent | | -0.34 | 0.01 | | -56.45 | 0.00 | | -0.35 | -0.33 |
| lender\_na | | -0.53 | 0.06 | | -8.66 | 0.00 | | -0.66 | -0.41 |
| lar\_count | | 0.00 | 0.00 | | -75.04 | 0.00 | | 0.00 | 0.00 |
| white\_cat2 | | 0.13 | 0.01 | | 18.58 | 0.00 | | 0.11 | 0.14 |
| white\_cat3 | | 0.25 | 0.01 | | 27.76 | 0.00 | | 0.23 | 0.26 |
| white\_cat4 | | 0.40 | 0.01 | | 35.63 | 0.00 | | 0.37 | 0.42 |
| metro\_90th | | -0.13 | 0.01 | | -14.67 | 0.00 | | -0.15 | -0.11 |
| metro\_80th | | -0.17 | 0.01 | | -16.21 | 0.00 | | -0.19 | -0.15 |
| metro\_70th | | -0.17 | 0.01 | | -14.17 | 0.00 | | -0.19 | -0.15 |
| metro\_60th | | -0.22 | 0.01 | | -15.07 | 0.00 | | -0.25 | -0.19 |
| metro\_50th | | -0.22 | 0.02 | | -13.22 | 0.00 | | -0.26 | -0.19 |
| metro\_40th | | -0.26 | 0.02 | | -13.36 | 0.00 | | -0.30 | -0.22 |
| metro\_30th | | -0.25 | 0.02 | | -11.96 | 0.00 | | -0.30 | -0.21 |
| metro\_20th | | -0.17 | 0.02 | | -7.18 | 0.00 | | -0.22 | -0.13 |
| metro\_10th | | -0.23 | 0.03 | | -8.54 | 0.00 | | -0.28 | -0.18 |
| metro\_less10th | | -0.22 | 0.03 | | -7.08 | 0.00 | | -0.28 | -0.16 |
| micro\_area | | 0.05 | 0.01 | | 3.51 | 0.00 | | 0.02 | 0.08 |
| metro\_none | | 0.24 | 0.02 | | 14.40 | 0.00 | | 0.20 | 0.27 |

The regression results were written into a new data frame and the standard error factor, coefficient and odds ratio are appended to it. The odds ratio will tell us how each independent variables has affected the outcome of loan denial.



Figure 26: Consolidation of results

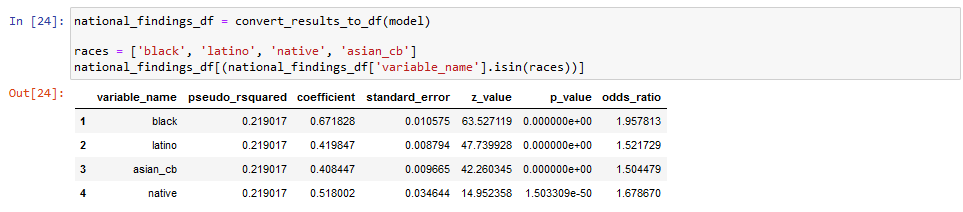


Figure 27: Odds ratios by race

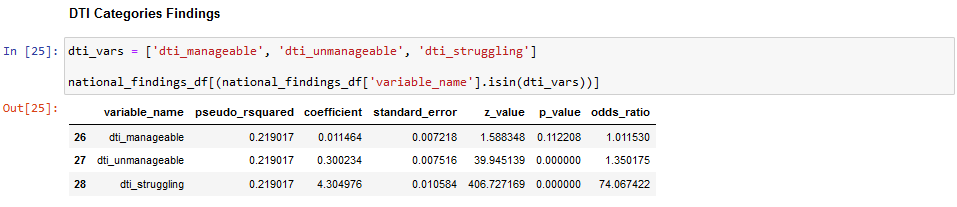


Figure 28: Odds ratio by debt-to-income ratio

An additional collinearity test was done to verify if the threshold for the variables was the same after the regression.

# ACCURACY OF THE STUDY

The performance of the calculation was verified using a confusion matrix as the dataset was imbalanced in terms of number of applicants per race, city and income levels.

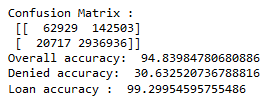


Figure : Confusion matrix

True negatives – 62929

False negative – 20717

False positives – 142503

False negatives – 2936936

# OBSERVATIONS AND RECOMMENDATIONS

# FINDINGS

The regression findings show that race of the applicant/co-applicant is a major factor in determining the outcome of the loan application with non-white applicants being more likely to be denied a loan even when they earn the same as their white counterparts, have a similar debt-to-income ratio and apply for the same amount of loan.

* Black applicants are almost twice as likely to be denied
* Latinx/Hispanic are almost 1.5 times
* Native Applicants are 1.7 times
* Asian/Pacific Islander are 1.5

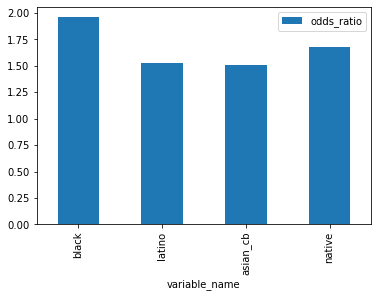


Figure 30: Graphical representation of odds by race

# RESULTS OF NYCKEL MODEL TRAINED WITH HMDA DATA

Nyckel is a cloud hosted machine learning platform that offers solutions to build, train, test and deploy machine learning models on cloud. The free developer tier allows only 1000 test samples to be submitted for training the system. The current dataset was reduced to 998 records in a prorated basis on the number of records per race and per outcome. This sample dataset was used to train the system and then the API was invoked using this data to test the results. The only value that changed each time the function was invoked was the race.

For the same set of individual variables and just the variance in race, the results show that all applicants other than white were rejected. Thus, showing that the data was biased and any ML model trained without human intervention is bound to be biased.

Table : Results of model trained on nyckel.com

|  |  |
| --- | --- |
| **Race** | **Outcome** |
| Native | Graphical user interface, text, application  Description automatically generated |
| Asian | A picture containing graphical user interface  Description automatically generated |
| Black | Graphical user interface, application  Description automatically generated |
| White | A picture containing graphical user interface  Description automatically generated |

# RECOMMENDATIONS

The logistic regression on the HMDA dataset for 2021 loan data in The United States of America shows that historical biases and the segregation against the minorities in the housing market among other social norms is also represented in the loan approval parameters.

The study has sufficiently proved that an AI models trained with such biased datasets also show a similar bias by considering race as an independent variable in determining the outcome of a loan application.

AI and ML are being used for a large number of applications ranging from applications that do not have any material impact on the society and people such as product recommendations, personalized ads and video recommendations to applications such as credit risk analysis, crime prediction and job recruitment systems.

Since the AI models are trained without human supervision, there is a chance for model to derive a correlation as a causation and determine a variable in the data as a decision factor for the outcome when in real world the same variable holds no weightage in the outcome. To minimize such mistakes, one should avoid using unsupervised machine learning models when the real-world stakes are high.

It is also worth noting that there will be additional challenges to accurately train AI models when the training data is coming from a data stream as it becomes more complicated for a human to identify any biases that are introduced in the later stages of the application lifecycle. Thus, periodic checks should also be conducted in such cases. One way to eliminate this significantly is to omit or remove the columns that are not relevant to the regression. E.g.: The data sent to a resume screening AI model should exclude all personal information such as age, gender, ethnicity and even addresses. It should only contain columns that represent data that a human recruitment consultant would look at such as years of professional experience and skills.

Lastly, it is important to include domain experts for the areas that are related to the application to ensure that they can validate and verify he accuracy of the model.

# SUMMARY

The rate of adoption of machine learning and cloud has been increasing multi-fold on in the Fintech sector. The literature review has suggested the industry wide usage of micro services and cloud computing as a base for the websites and mobile apps used by various organizations to their customers. It also used for fraud detection, application security, marketing, sales and customer service.

Since the repercussions of a faulty AI/ML model in financial institutions may have adverse impact on the lives of customers/users, a great deal of caution and attention to detail are required during the development and deployment of such applications. AI/ML can add a lot of value to businesses as long as the algorithms are accountable and explainable.

# REFERNCES

Arango Kure, M. (2022) “Brand Equity and Data Privacy: Beyond Selling Eyeballs: A business case for online privacy-friendly data practices in regions covered by the GDPR.”

Arslanian, H. and Fischer, F. (2019) *The future of finance: The impact of FinTech, AI, and crypto on financial services*. Springer.

Awoyemi, J.O., Adetunmbi, A.O. and Oluwadare, S.A. (2017) “Credit card fraud detection using machine learning techniques: A comparative analysis,” in *2017 International Conference on Computing Networking and Informatics (ICCNI)*, pp. 1–9. https://doi.org/10.1109/ICCNI.2017.8123782.

Bao, W., Lianju, N. and Yue, K. (2019) “Integration of unsupervised and supervised machine learning algorithms for credit risk assessment,” *Expert Systems with Applications*, 128, pp. 301–315. https://doi.org/https://doi.org/10.1016/j.eswa.2019.02.033.

Christine Ro (2021) *Helping workers understand their implicit biases is helpful, in theory – but these programmes are controversial. Who’s right?* *BBC*.

*Consumer Response Annual Report* (no date).

Daumeyer, N.M. *et al.* (2019) “Consequences of attributing discrimination to implicit vs. explicit bias,” *Journal of Experimental Social Psychology*, 84, p. 103812. https://doi.org/https://doi.org/10.1016/j.jesp.2019.04.010.

Ferrer, X. *et al.* (2021) “Bias and Discrimination in AI: A Cross-Disciplinary Perspective,” *IEEE Technology and Society Magazine*, 40(2), pp. 72–80. https://doi.org/10.1109/MTS.2021.3056293.

Freiberg, M. (2022) “Third-Party-Tracking bei Wiley und Springer: Analyse und Ausblick,” *ABI Technik*, 42(2), pp. 96–104. https://doi.org/doi:10.1515/abitech-2022-0017.

Goodyear, M.P. (2022) “Returning to the Start? Federal BIPA Claims after Transunion v. Ramirez,” *U. Ill. L. Rev. Online*, p. 10.

Gotthardt, M. *et al.* (2020) “Current state and challenges in the implementation of smart robotic process automation in accounting and auditing,” *ACRN Journal of Finance and Risk Perspectives*, 9(1), pp. 90–102. https://doi.org/10.35944/JOFRP.2020.9.1.007.

Gotze, M. *et al.* (2022) “Measuring Web Cookies in Governmental Websites,” in *14th ACM Web Science Conference 2022*. New York, NY, USA: ACM, pp. 44–54. https://doi.org/10.1145/3501247.3531545.

Hoikkala, I.-M. and Ojala, R. (2022) “Customer Engagement Through the Use of AI: A Qualitative Study on SMEs and The Use of Chatbots.”

Howard, A. and Borenstein, J. (2018) “The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity,” *Science and Engineering Ethics*, 24(5), pp. 1521–1536. https://doi.org/10.1007/s11948-017-9975-2.

https://www.fico.com/en/newsroom/housing-bubble-inflating-mortgage-lenders-tell-fico (2014) *Housing Bubble Inflating, Mortgage Lenders tell FICO*.

Huang, J.Y., Gupta, A. and Youn, M. (2021) “Survey of EU ethical guidelines for commercial AI: case studies in financial services,” *AI and Ethics*, 1(4), pp. 569–577.

Huttunen, J. *et al.* (2019) *BIG DATA, CLOUD COMPUTING AND DATA SCIENCE APPLICATIONS IN FINANCE AND ACCOUNTING*, *ACRN Journal of Finance and Risk Perspectives*. Available at: https://www.datamation.com/big-data/structured-vs-unstructured-data.html.

Kamal, Y. (2016) “Study of Trend in Digital Marketing and Evolution of Digital Marketing Strategies,” *International Journal of Engineering Science and Computing*, p. 5300. https://doi.org/10.4010/2016.1298.

Louis Columbus (2020) *The State Of AI Adoption In Financial Services*, *Forbes*.

Luguri, J. and Strahilevitz, L.J. (2021) “Shining a Light on Dark Patterns,” *Journal of Legal Analysis*, 13(1), pp. 43–109. https://doi.org/10.1093/jla/laaa006.

Madhusudhan, R. and Surashe, S. v (2022) “Privacy and Security Comparison of Web Browsers: A Review,” in L. Barolli, F. Hussain, and T. Enokido (eds) *Advanced Information Networking and Applications*. Cham: Springer International Publishing, pp. 459–470.

Mehrabi, N. *et al.* (2021) “A Survey on Bias and Fairness in Machine Learning,” *ACM Comput. Surv.*, 54(6). https://doi.org/10.1145/3457607.

Mei, L. (2022) *Fintech Fundamentals: Big Data/Cloud Computing/Digital Economy*. Stylus Publishing, LLC.

Mogaji, E., Soetan, T.O. and Kieu, T.A. (2020) “The implications of artificial intelligence on the digital marketing of financial services to vulnerable customers,” *Australasian Marketing Journal*, p. j-ausmj.

Nguyen, D.K., Sermpinis, G. and Stasinakis, C. (2022) “Big Data, Artificial Intelligence, and Machine Learning: A Transformative Symbiosis in Favour of Financial Technology,” *European Financial Management* [Preprint]. https://doi.org/10.1111/eufm.12365.

Nguyen, V.H.A. (2022) “An Agile Approach for Managing Microservices-Based Software Development: Case Study in FinTech,” in M. Themistocleous and M. Papadaki (eds) *Information Systems*. Cham: Springer International Publishing, pp. 723–736.

Ntoutsi, E. *et al.* (2020) “Bias in data-driven artificial intelligence systems—An introductory survey,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3). https://doi.org/10.1002/widm.1356.

Pantielieieva, N. *et al.* (2018) “FinTech, Transformation of Financial Intermediation and Financial Stability,” in *2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T)*, pp. 553–559. https://doi.org/10.1109/INFOCOMMST.2018.8632068.

Qi, Y. and Xiao, J. (2018) “Fintech: AI powers financial services to improve people’s lives,” *Communications of the ACM*, 61(11), pp. 65–69.

Riva, A. and Pilotti, L. (no date) “One-to-one Marketing. How to Build an Effective Digital Marketing Strategy.” https://doi.org/10.13132/2038-5498/12.3.329-349.

Rubel, A., Castro, C. and Pham, A. (no date) *Democratic Obligations and Technological Threats to Legitimacy This material is forthcoming in revised form in Algorithms & Autonomy: The Ethics of Automated Decision Systems*. Available at: https://web.archive.org/web/20191216105955/https://dailybruin.com/2013/04/26/professor-helps-develop-.

*The Digital Future of Finance Gartner for Finance* (2021).

*The IT Roadmap for Digital Business Transformation Excerpt* (2021).

*The Markup* (2021).

Tonya R. Moon (2011) *Project Implicit Test*, *https://implicit.harvard.edu/implicit/takeatest.html*.

Waldman, A.E. (2020) “Cognitive biases, dark patterns, and the ‘privacy paradox,’” *Current Opinion in Psychology*, 31, pp. 105–109. https://doi.org/https://doi.org/10.1016/j.copsyc.2019.08.025.

Angwin, J., 2022. *Confronting the Biases Embedded in Artificial Intelligence.* [Online]   
Available at: https://themarkup.org/newsletter/hello-world/confronting-the-biases-embedded-in-artificial-intelligence

Bhattacharya, S., 2021. *Basic Steps in Machine Learning-Based Modeling.* s.l.:s.n.

Columbus, L., 2020. *The State Of AI Adoption In Financial Services.* [Online]   
Available at: https://www.forbes.com/sites/louiscolumbus/2020/10/31/the-state-of-ai-adoption-in-financial-services/?sh=5014caac2aac

Dastin, J., 2018. *Reuters.* [Online]   
Available at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Emmanuel Martinez, L. K., 2021. *The Secret Bias Hidden in Mortgage-Approval Algorithms.* [Online]   
Available at: https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms

Jeschke, R., 2018. *Affordances: Science Fiction About Algorithmic Bias and Technological Resistance.* [Online]   
Available at: https://www.eff.org/deeplinks/2019/11/affordances-science-fiction-about-algorithmic-bias-and-technological-resistance  
[Accessed ].

Martinez, E. & Carollo, M., 2021. [Online]   
Available at: https://themarkup.org/denied/2021/10/22/markup-investigation-cited-by-officials-in-announcing-new-mortgage-discrimination-enforcement

OpenAI.org, 2022. [Online]   
Available at: https://github.com/openai/dalle-2-preview/blob/main/system-card.md?utm\_source=Sailthru&utm\_medium=email&utm\_campaign=Future%20Perfect%204-12-22&utm\_term=Future%20Perfect#bias-and-representation

PWC, 2021. *AI Predictions 2021.* [Online]   
Available at: https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-predictions.html

# APPENDIX A: RESEARCH PLAN

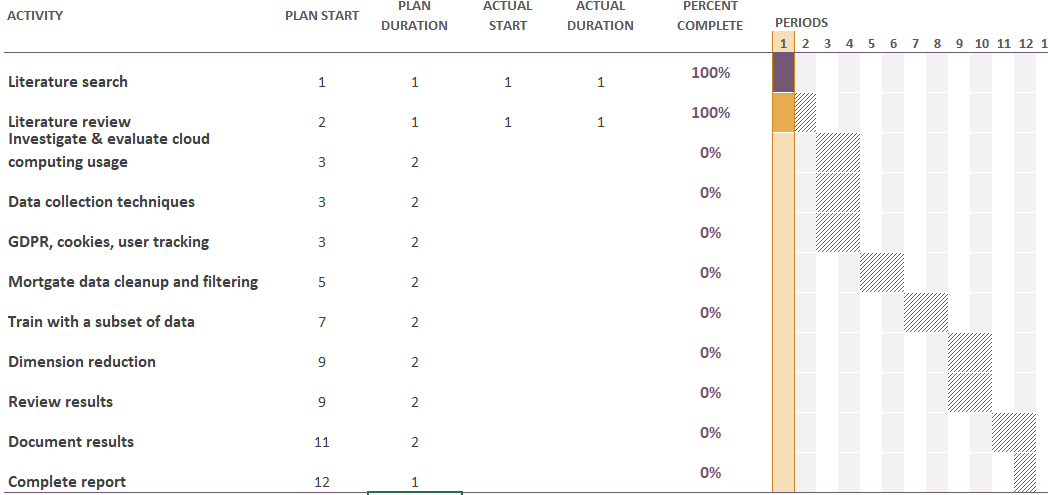


Figure 31 : Research plan

# APPENDIX B: RESEARCH PROPOSAL

This research aims to study the different ways Cloud computing and Machine learning technologies are used in companies providing financial services. It also looks for possible regulatory and discriminatory processes in the system and research ways to identify and resolve any such practices introduced into the system because of technology.

# BACKGROUND

Though the future looks bright for the applications of AI and ML in most of the industry sector, the biases(Howard and Borenstein, 2018) and the resulting causalities of poorly designed and implemented Algorithms is becoming more common. This is because the regulations and laws on AI are non-existent and most of the implementations are opaque. Even companies well renowned for their technical competence have been unsuccessful in some of their attempts to use Machine learning to take decisions for them. The most infamous case being Amazon’s AMZN.O algorithm for screening resumes showing a bias against women. (Dastin, 2018).

When machine learning is poorly implemented by streaming services to recommend videos or by e-commerce/retail industries to create custom offers for customers, the real-world impact of this is negligible. But when it comes to financial services where small mistakes can lead to Individuals suffering and large mistakes can even lead to economic crashes. With fraud detection and credit screening(Louis Columbus, 2020) being the most used applications of AI, the risk of Machine learning algorithms discriminating against particular groups in the society is also a real possibility.

Research by a non-profit organization called theMarkup on the biases in Mortgage Approval Algorithms used in the United States (Emmanuel Martinez, 2021) has sparked the debate about Algorithmic Accountability. (Martinez & Carollo, 2021).

Good machine learning algorithms need clean data and in ideal cases multiple point of truths. But with regulations over data collection, handing and processing non-existing outside of the European Union’s GDPR law, also with various functions of these financial originations being outsourced or using SAAS and PAAS offerings available in the market, these algorithms are becoming black boxed (Angwin, 2022)

The real-world biases are often introduced into the data lead to the AI algorithms also exhibiting these biases. (OpenAI.org, 2022)

# PROBLEM STATEMENT

There will always be biases in the society and thus the probability of the same biases being introduced into the data is also very high(Ntoutsi *et al.*, 2020; Ferrer *et al.*, 2021). Since, machine learning algorithms are isolated in terms of their perception of the world, they are by design unable to differentiate between correlation and causation.(Mogaji, Soetan and Kieu, 2020) In other words, they do not have common sense. The only world the data that is fed to them. So, they never identify or fill the gaps. (Bhattacharya, 2021)

For example, the case of Amazon resume screening algorithm discriminating against female applicants was because the data used to train the algorithm on suitable profiles was based on the resumes of the existing Amazon staff and the lack of diversity that was pre-existing in the organization was seen as women being undesirable for employment by the AI algorithm. (Dastin, 2018).

Cloud computing and Machine learning are being used for entire customer journey from marketing, lead generation, customer onboarding, services, support and upselling/cross selling(Huttunen *et al.*, 2019).

There is a need to look into how these emerging technologies are being used, existing ethical issues, impact of this on the customer and ways to mitigate these.(Gotthardt *et al.*, 2020)

# RESEARCH QUESTIONS

The following research questions are suggested for each of the research objective as highlighted as follows.

1. What are the different sub domains in financial services where Cloud computing, AI, ML are used.?
2. What are the different cloud based offerings available for Financial services?
3. How to identity and eliminate biases in Machine learning?

# AIM AND OBJECTIVES

The primary objective of the current research is to study the way technology is being adopted in Fin Serv and propose a method to identify techniques to eliminate biases

The study purposes are formulated as follows:

* To analyse the different areas of business where cloud computing and AI are being used
* To compare how AI SAAS offerings such as Nyckel, Lobe compare to traditional data science approach
* To suggest ways to identify and predict potential issues in data

# SIGNIFICANCE OF THE STUDY

Unlike a decade ago, when technology was mostly limited to people who where literate in computer science, the affordability of smart phones and the internet has lead to people from all walks of like becoming exposed to technology. The entry barrier to use the present day computing machines is very low. Thus, most people who use an app or a website for example do not know how they work, so they cannot be expected to act in their own interests. So, there is a need for research and push for regulation on how technology is used especially in sectors that can make or break people’s lives.

A rejected mortgage or a loan can have a devastating impact on a person and their family. When decisions of such high importance are being delegated to computers, it is important to analyse and measure the accuracy of these decisions.

# SCOPE OF THE STUDY

The study will comprise the following:

1. What services offered by the Financial services organizations to their customers use Cloud computing.
2. The different data collection techniques used by financial services industries
3. Analyse the Dynamic National Loan-Level Dataset available on FFIEC website to look identify discrimination patterns and find ways to solve them

# RESEARCH METHODOLOGY

The loan application data and Transmittal sheet records are available for United states mortgage approval and rejection data for the years 2017 through 2020.

The plan is to filter the data to first-lien mortgages for home purchase, where the borrower intends to live in the property. Then to further filter, using only clearest outcome: loans either approved or denied. All other outcomes, such as applications that are withdrawn and applications that are approved but not accepted by the applicant are excluded.

There are about 15 variables or dimensions identified in the data

Table 7: Proposed variables for analysis

|  |  |
| --- | --- |
| 1 | Age |
| 2 | Combined loan-to-value ratio |
| 3 | Credit model used |
| 4 | Debt-to-income ratio |
| 5 | Income |
| 6 | Loan amount |
| 7 | Location/City and size |
| 8 | Mortgage term |
| 9 | Property value |
| 10 | Race |
| 11 | Sex |
| 12 | The automated underwriting system used |
| 13 | The size of the lender |
| 14 | The type of lender |
| 15 | Whether the application had a co-applicant |

Will be leveraging sklearn libraries to use logistic regression to train the model with existing data and then use combinations of test data to determine which of the dimensions had higher weightage on outcome.

Further will train an algorithm with a subset of the above dimensions to validate if there is a change in the outcome and if dimension reduction techniques would be useful on financial datasets.

# RESOURCE REQUIREMENTS

* Anaconda 3 or later
* Since Python stores all data frames in memory. The RAM of the machine used for data cleanup, transformation and analysis should be at least 128 gigabytes.

# APPENDIX C: ETHICS FORMS

The study doesn’t involve any participants and all the data used for the research is available in the public domain and doesn’t have any identifiers to any individuals.

I also here by confirm that I do not have any conflict of interest with respect to the technologies and products mentioned in this study.