**Module 8 – Critical Thinking Assignment**

Lucas Rutledge

Colorado State University Global

MIS581 Capstone: Business Intelligence and Data Analytics

Dr. Steve Chung

5/11/2023

**MIS581 Module 7 – Portfolio Project**

**ABSTRACT** – Due in Week Seven

This will be on a separate page. Make sure you have a page break between the abstract and the rest of the paper.

**INTRODUCTION**

Every industry has the potential to benefit from data research and analysis in some way or form. For this first module, our goal is to identify one industry/organization that could see benefit from a research proposal. One of the largest industries that has yet to fully capitalize on the benefits of data analysis is the financial and banking industry. Banking has become almost second nature in current society with pretty much all forms of saving, spending, and borrowing money being conducted through a bank account. However, there have been many risks identified over the years to the successful long-term implementation of analytics in banking. While finding the right way to implement analytics into banking for a long term can be challenging, there is still a good amount of success that can be obtained through increased analytics. Vivek (2022) explores the ways that financial institutions have currently been using analytics in their business models. The four listed areas come from managing fraud detection to account for risk, modelling market risk to determine and forecast the institutions standing, analyzing credit risk to determine borrower strength, and operational risk to ensure operations are maintaining at a sustainable level. Each of these observed uses can be a challenge for an institution that has not yet been able to capitalize. Garg et. al. (2017) found that 54 percent of the top 500 financial institutions in the industry are priced below book value, and that just 18 percent of those banks capture virtually all the value. While all of these can be areas of research, the problem I will address will be to optimize the process of the underwriting team to make decisions on the credit worthiness of an applicant through data analysis.

**OBJECTIVES**

The aim of this study is to assist in optimizing the decision-making process for the lending department to make decisions on the credit worthiness of an applicant through data analysis. In order to do this, I will be using a sample dataset of existing loan applications which contains information about the applicant and the approval status. Next the data must be processed and tested to determine any undesirable trends for the decisions made for previous applications. With any level of discriminate practice being found, a proposal will be made on how to eliminate that practice. This will then pave way for a baseline standard of loan eligibility to be followed at the institution level. The goal of this proposal will be to create a baseline system to determine loan eligibility that is impartial to the consumer and provide security for the institution that’s using it. As well, we can validate existing loan practices and verify if there is any sense of discrimination or unfair treatment of the customer base for an institution. Boatwright and Sullivan (2017) explain that discriminatory practices that are overt or have a disparate treatment of the consumer or impact to them can have very negative impacts on your fair lending programs. These fair lending practices are upheld through numerous government regulations. Our goal is also to increase the speed at which a decision can be made, allowing for the opportunity of business growth and customer retention.

**OVERVIEW OF STUDY**

The banking industry provides a perfect foundation for the incorporation of data analytics. According to the FDIC in 2020, nearly 95 percent of households in the United States have a bank or credit union account. Furthermore, SelectUSA (2020) shares that banking is one of the largest sectors of the economy, as finance and insurance services represented 8.3 percent of the United States gross domestic product. Financial institutions have shown a large impact on the daily lives of many households both in personal finance and market sustainability. Because of how large the consumer base is and how often finances are utilized in households, there is a very high volume of data points that are generated everyday pertaining to so many different elements of the consumers livelihood. Their spending habits and frequency, fraud risk, credit worthiness, and loyalty to a specific institution can all be found by some level of analysis. However, the banking industry has been very slow to adopt a sustainable analytic model. Garg et. al (2017) found that while over 90 percent of top banks have used advanced levels of analytics, most of them have only found singular projects of success and no continuation to scale up. Despite this, most of these banks also state that they find the use of data analytics would provide a competitive advantage for them within the industry market. Shalimov (2019) shares a few challenges banks face with successfully implementing analytics and explains that 92 of the top banks have not updated their main operating systems used for decades at this point. When it comes to processing big sources of data there’s an increased risk of the system failing resulting in greater privacy risks. Both these challenges require a large financial investment in the resources required to maintain security, which can be hard to generate buy in from leadership which would uphold the necessary resources.

**RESEARCH QUESTIONS AND HYPOTHESES**

Research questions are a foundation that help pinpoint the issue that has been brought up or identified. Ang (2014) goes into great detail about research questions and how they can be formed over the initiation of a new project. To begin, we look at the source of the research questions which in my proposal is a practice-oriented source. Ang shares that this is driven by the need of a specific organization or industry, where the results may be specific to just that area and have no generalized use. From there, the methodology is considered, and the data is collected. The last note from Ang I see is that there is a lot of subjectivity and refinement that can come from the creation of business questions. If you are new to an industry, your understanding of what makes for a good business question for that industry will take time. From all this information, I have two questions that I will begin my research with.

* Question 1 - Is the sample financial institution treating their loan applicants fairly? Specifically, this is looking to identify if there are any qualitative variables that have a significant impact on the approval of a loan.
* Question 2 - Does the income an applicant has or the amount the applicant is applying for have any significant impact on the application approval? This then looks quantitatively at different numerical inputs that may influence the loan to find trends in approvals.

Now that the research questions have been identified, a hypothesis set must be established for each. Rawat (2021) explains that hypotheses are the verification of assumptions that will answer the research questions. Rawat states that both the null and alternative are treated as equally important for testing, however the alternative is the driving force of our research question. For the two questions identified, each will have a set of two hypotheses, one null stating no relationship, and an alternative.

* Question 1
  + H0: There is no significant relationship between gender, marital status, education level or property area and the chance for loan approval.
  + HA: There is a relationship between any gender, marital status, education level or property area and the chance for loan approval.
* Question 2
  + H0: There is no significant relationship between either applicant income nor amount applied for and the chance for loan approval.
  + HA: There is a relationship between either applicant income and/or amount applied for and the chance for loan approval.

**LITERATURE REVIEW**

The first research article chosen comes from Doko and Mishkovski in 2019. They talk about the advantages and disadvantages banks are currently faced with incorporating big data into their business model. This paper is relevant to my research in two sections, it’s information on the risks of big data and its context example of credit risk assessment. Doko and Mishkovski use qualitative research to study the types of infrastructure banks utilize to store their information. They state that “Because of strict governance rules in banking, banks are missing special working positions for data scientists. Also because of their standards and regulations banks are behind the other technologies in analytical capabilities…” (Doko and Mishkovski, pg. 13). As well, Doko and Mishkovski explain that without the proper skilled workers to synthesize the data and answer areas of research, there is no benefit to processing so much data.

As well, Doko and Mishkovski (2019) share how this research can be incorporated to risk-based assessment of credit. This supports my own project trying to establish a method of risk assessment to streamline the lending process.

The second research article is from He et. al. in 2022 by looking at IT spending in banking. He et. al. review a large sample of large and small banks across the U.S. to find trends of spending on technology in different areas of business, including lending. They utilize quantitative analysis and conduct a regression on the spending of banks in the IT sector to find a correlation between spending in the industry and use of big data. This is relevant to my research as the ability to improve data analytics and business intelligence comes from updated tech to provide extra security and efficiency. Understanding how much institutions are willing to (or even are able to) spend on modernizing their tech is key to understanding potential limitations to incorporating the research I’m conducting.

The third article to support my research is by Chavan and Gambacorta (2016). They look at sample information from India to review the issue of procyclicality and non-performing loans, or NPLs. Procyclicality is the trending of financial variables to fluctuate during economic cycles, which overall can provide false or fleeting outlooks on individuals financial strength and overall risk. Using regression as well, Chavan and Gambacorta uncover that “When the NPL ratio of Indian banks is juxtaposed with bank credit growth graphically, a lagged positive association is broadly discernible.” (Section 4). This review shares that financial institutions use herd behavior to lending over time in order to compete for business, without looking at inherent risks on a case by case basis. This is due to institutions being independent for the most part, but “driven by two short-term concerns: earnings and reputation.” (Chavan and Gambacorta, Section 2). As well, these financial institutions also can have fleeting memory of prior credit-busts and the tendencies that can become risky.

Bartlett et. al. (2019) is the final area of research I have added, going over the discrimination that can occur in lending. The purpose of their study is to estimate discrimination levels in the home lending sector of bank lending. To do this, Bartlett et. al use a myriad of different applicant groups to find quantitative trends in their approval odds from an automated loan processor. They found that “the interpretation of these main-credit-risk-model results is that conditional on being given a loan, African-American and Latinx borrowers pay an average of 7.92 basis points more than other borrowers for their purchase mortgage.” This is pertinent information to my own research, as we have seen how heavily regulated the industry is revolving around avoiding loan discrimination, including the Equal Credit Opportunity Act of 1974, Truth in Lending Act of 1968, and Fair Housing Act of 1968.

**RESEARCH DESIGN**

**Methodology**

The chosen dataset takes information about applicants for several loan applications and the approvals of each. The dataset has 13 variables and over 600 rows with each representing an application. Of all the variables, six of them are considered binary in nature meaning that they only have at most two possible outputs. These six are Gender, which looks at the gender of the applicant, Married which determines marital status of the applicant, Education which determines if the applicant is a graduate or not, Self-employed which determines if the applicant has their own business or earns income from an employer, Credit History to determine if the applicant has a history of credit at all, and Loan Status which looks at if the applicant was approved or not for the loan. There are also five variables that are interval, looking mainly at a scale across numerical values. These include Applicant income, Co-applicant income, the Loan amount and the loan term. Lastly there are two text variables with the application number and the Property location.

Figure 1: Dataset Dictionary

Table

Description automatically generated

**Methods**

In order to properly analyze the dataset, tools will need to be utilized to process and configure the dataset, and models will need to be made to test it. First, the use of descriptive statistics and frequency analysis will be beneficial to providing context to the dataset and further explore the research questions. From there, different types of testing will be needed for the quantitative and qualitative data. According to a report of statistical analysis types from UCLA, the best way to research the first question is through a Chi-square test. With a Chi-square, we are comparing the two groups against each other for each combination to determine our results. Meanwhile, for the second question, we are working with two numerical values and one categorical variable. The question is to identify if any relationship exists between each numerical value to our categorical. The best way to test for significance for the hypothesis is through Kruskal-Wallis test. UCLA explains this is a non-parametric version of ANOVA, and is used primarily to compare categorical independent variables with at least two groups, and an ordinal numerical dependent variable.

**Limitations**

The limitations of this study are largely based in the granularity of the applicant information obtained in the dataset. For the first question regarding lending standardization, the only two variables available to test are applicant income and loan amount. According to the Bank of America website (2023), there are more factors that go into the approval process. As well as the amount requested and the ability to repay, there are also the conditions that you need the loan for and your intent for taking out the money. On top of this, credit history is vital to understanding loan stability and projected risk on taking new credit. These all make up what Bank of America states are the 6 C’s of creditworthiness. For the second question, while we can test for discrimination for gender and education status, this also lacks further testing to racial disparity. Having more variables to test would provide more ways to test the hypotheses and therefore further ability to validate the questions.

**Ethical Considerations**

This dataset takes sensitive information from customers of a financial institution. In order to maintain privacy and security of those sampled, the names of the applicants must be omitted. On top of names, there are several other personal identifying pieces of information that can be gathered for the application of a loan that can jeopardize consumer privacy. Private contact information including phone numbers and addresses, names of employers, and social security numbers are all present on accounts. These all need to be removed before any information about the data can be opened to test in any capacity.

Ensuring privacy in data testing is one of the most scrutinized areas of banking. Failure to keep private information secure can have a multitude of negative effects including distrust by the customers, fines, and other punishments from the federal regulators. This can prove costly to the organization, as can failing to follow the regulations of fair lending and consumer equality. It is important that these findings identify any potential failures so action plans can be made before they cause issues.

**FINDINGS (Due in Week Six)**

The two questions chosen will utilize two numerical variables and four categorical variables to test against. To better understand the dataset, means testing and frequency analysis will provide an overview of the test statistics.

Figure 2: Statistic and Frequency Analysis

A screenshot of a computer

Description automatically generated with medium confidence

The testing of question one involves comparing four qualitative variables to a different qualitative variable. These are not able to be tested as numerical trends, rather the use of a comparative analysis is required to determine the possibility of a relationship. A report of statistical analysis types from UCLA shares that the best way to look for this is through a Chi-square test. With Chi-square, two categorical groups are assumed to be mutually exclusive and the null hypothesis is that these classes remain independent. When the test is conducted, the values from each are tested assuming the null to determine the frequency in which it is followed. Conducting the test for each of the four variables produced the following results.

Figure 3,4,5,6: Chi-square test for Gender, Married, Education, and Property Area

A screenshot of a computer

Description automatically generated with low confidenceA screenshot of a computer

Description automatically generated with low confidenceA screenshot of a computer screen

Description automatically generated with low confidenceA screenshot of a computer screen

Description automatically generated with low confidence

To interpret the results, we observe the p-value as it relates to the significance variable, with a usual level of confidence 0.05. The chi-square tells us how many degrees of freedom each variable has, which correlates to “the number of independent variables that can be estimated in a statistical analysis…” (Ganti, 2023). Gender, Married, and Education have one degree of freedom, while Property Area has two. This signifies that the first three can have at most one independent estimate before we must constrain the data to prevent including all the values, while the last can go up to two before it needs to be constrained. When it comes to Gender, we have a p-value=0.6264, this is greater than our usual significance level of 0.05 and therefore we fail to reject the null hypothesis, meaning there is not sufficient evidence to suggest a relationship between gender and loan approval. For the Married variable, we have a p-value=0.0273 which is less than our usual significant level. Therefore, we reject the null hypothesis for marital status, there is sufficient evidence to suggest a relationship between marital status and loan approval. For the Education level there is a p-value=0.033 which again is less than our usual significance level. Again, we can reject the null hypothesis, there is sufficient evidence to suggest a relationship between education status and loan approval. Lastly, for the Property Area we have a p-value=0.0021 which is also less than the significance level. This again means we can reject the null hypothesis for property area, there is sufficient evidence to suggest a relationship between residence area and loan approval.

For question two, we are testing a numerical value to a categorical variable in two different cases. To test for significance, the Kruskal-Wallis method will be used to compare the means of each numerical variable to the two categorical groups. The groups here would be applicants whose loans were approved, and applicants whose loans were not approved. The Kruskal-Wallis method assumes that the interval variable is not normally distributed.

Figure 7,8: Kruskal-Wallis test for ApplicantIncome and LoanAmount

A screenshot of a computer

Description automatically generated with low confidenceA screenshot of a computer

Description automatically generated with low confidence

Beginning with Applicant Income, there is a p-value=0.9179 with one degree of freedom. This is greater than the usual significance level of 0.05, meaning we fail to reject the null hypothesis. There is not sufficient evidence of a relationship between the income for the applicant and the approval of a loan application.

Figure 9: Distribution of Scores for Applicant Income

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

For the Loan Amount, the p-value=0.3951 which is again greater than our usual significance level. This means we fail to reject the null hypothesis as well, there is not sufficient evidence of a relationship between the loan amount applied for and the approval.

Figure 10: Distribution of Scores for Loan Amount

A picture containing text, screenshot, diagram, rectangle

Description automatically generated

**CONCLUSION**

Overall, of the six tests conducted, three produced sufficient evidence to support the alternative to the null hypothesis while three failed to find anything significant to object the null. Beginning with question one, we observed that there was no statistical evidence that Gender would have a relationship with the loan approval. This does not mean that a relationship can’t exist, however our data did not indicate this. However, the Married, Education, and Property Area did all indicate some evidence that a relationship could be inferred to the loan approval. Based on these findings, this means that the decision making of the loan underwriters for this sample data could factor applicants marital status, education level and residence into their approval or denial of a loan. This is cause to raise concern to the leadership that fair lending practices may be at risk.

For our second question, we found that there was no statistical evidence that neither Applicant Income nor LoanAmount were significant to the decision of a loan approval. This means that we do not have any statistically significant information to establish a lending tool to determine odds of approval with the data we have currently.

**RECOMMENDATIONS**

The organization should further investigate their lending practices to find other possible areas of discrimination in application decisions. To ensure adherence to fair lending and government regulations, the root cause of why areas like marital status, education, and residential area can be related to the applicants ability to be approved for a loan. From there, leadership directive should provide training and monitoring for future loans to prevent this from happening further.

While I believe that there was not anything available from this research to direct a lending process improvement, this does not mean that the research was a failure. The data collected did not include as many variables related to a loan application as would be needed to properly ensure a tool was complete. The testing suggests that income and loan amount are not important relative to the decision of the loan, this still provides benefit in that they would not be useful in what the organization would need. More resource gathering could collect more information from a loan application, such as credit score, credit history, and loan purpose. Additional testing on these variables could factor other elements of the 6 C’s of creditworthiness into a more inclusive tool for the organization moving forward.

**REFERENCES**

Ang, S. H. (2014). *Research Design for Business & Management*. <https://doi.org/10.4135/9781473909694>

*Bank of America.* Bank of America - Banking, Credit Cards, Loans and Merrill Investing. (2023, April 3). Retrieved May 7, 2023, from https://www.bankofamerica.com/smallbusiness/resources/post/factors-that-impact-loan-decisions-and-how-to-increase-your-approval-odds/

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2019). *Consumer-lending discrimination in the Fintech Era.* FDIC*.* <https://doi.org/10.3386/w25943>

Boatwright, K., & Sullivan, K. (2017, June 20). What is a real-world example of fair lending discrimination? *Ncontracts.* Retrieved March 23, 2023, from <https://www.ncontracts.com/nsight-blog/3-real-world-examples-fair-lending-discrimination>

Chavan, P., & Gambacorta, L. (2016). Bank lending and loan quality: the case of India . *BIS Working Papers*, 595.

Doko, Fisnik & Mishkovski, Igor. (2019). An Overview of Big Data Analytics in Banking: Approaches, Challenges and Issues. *UBT International Conference. 270.* <https://www.researchgate.net/publication/352362924_An_Overview_of_Big_Data_Analytics_in_Banking_Approaches_Challenges_and_Issues>

Ganti, A. (2023, April 3). *Degrees of freedom in statistics explained: Formula and example*. Investopedia. Retrieved May 7, 2023, from <https://www.investopedia.com/terms/d/degrees-of-freedom.asp>

Garg et al. (2017, June 22). *Analytics in Banking: Time to realize the value*. McKinsey & Company. Retrieved March 26, 2023, from <https://www.mckinsey.com/industries/financial-services/our-insights/analytics-in-banking-time-to-realize-the-value>.

He, Z., Jiang, S., Xu, D., &amp; Yin, X. (2022). Investing in bank lending technology: IT spending in banking. *NBER Working Paper Series*. <https://doi.org/10.3386/w30403>

*How america banks: Household use of banking and Financial Services, 2019 FDIC survey*. (2020, October 19). FDIC. Retrieved March 25, 2023, from <https://www.fdic.gov/analysis/household-survey/index.html#:~:text=The%202019%20survey%20finds%20that,was%20first%20conducted%20in%202009>.

*Positivism.* StudySmarter US. (n.d.). Retrieved April 16, 2023, from <https://www.studysmarter.us/explanations/social-studies/theories-and-methods/positivism/#:~:text=The%20positivist%20approach%20to%20research%20methods,-Researchers%20who%20adopt&amp;text=The%20goal%20of%20positivist%20research,best%20done%20through%20quantitative%20methods>.

Rawat, S. (2021, July 22). *What is hypothesis testing? types and methods.* Analytics Steps. Retrieved April 16, 2023, from <https://www.analyticssteps.com/blogs/what-hypothesis-testing-types-and-methods>

*SelectUSA Financial Services Industry*. International Trade Administration | Trade.gov. (2020). Retrieved March 26, 2023, from <https://www.trade.gov/selectusa-financial-services-industry>

Shalimov, A. (2019, January 19). *Big Data in the banking industry: The main challenges and use cases: Eastern peak.* Eastern Peak - Technology Consulting & Development Company. Retrieved March 26, 2023, from <https://easternpeak.com/blog/big-data-in-the-banking-industry-the-main-challenges-and-use-cases/>

Vivek, J. (2022, August 22). *Data Analytics in banking and Financial Services*. Zuci Systems. Retrieved March 26, 2023, from <https://www.zucisystems.com/blog/how-is-data-analytics-used-in-finance-and-banking-sector/>