Team Name: Los Datos

Student Loan Prediction and Repayment Risk

What are the potential risks associated with student debt and what are the best methods of mitigation?

Team Website: teamlosdatos.wixsite.com/studentloan
Team Email: teamlosdatos@gmail.com
Dataset: https://www.brookings.edu/research/the-student-loan-crisis-a-look-at-the-data

Authors:
Alfredo Antolinez
Misti Stevens
Omar Waller

1. INTRODUCTION

Is college education the best investment after high school? How much will that investment cost? If student loan debt is inevitable, how much can I expect to pay given the institution I attend? We hope to answer these questions and more throughout this project by performing a detailed analysis on a dataset used by Adam Looney at Brookings, a nonprofit public policy organization based in Washington, DC.

Issues with student loans and financial aid touch nearly every college applicant across the nation. Tuition has increased exponentially over the last 50 years. This project is interesting because we hope to predict several key indicators like Overall Outstanding Principal Balance and Overall Repayment Rate for schools provided in the dataset. Additionally, we hope to predict student loan debt given indicators like the parent's adjusted gross income (AGI) or an independent student's AGI when applying to college.

The beneficiaries of this project will be any college applicant looking to find information about student loan data given a school. This project will also give applicants an idea of what may be expected, given their financial makeup, regarding student loans at an institution.

1.1 Project goals

The primary objectives of this project are to...

- Classify the risk of borrowing student loans at various post-secondary institutions
- Determine predictive reasoning capabilities of provided attribute data for determining the likelihood of student loan repayment
- Create a model to predict the loan repayment risk associated with attending a particular school

Each of these tasks will be expanded upon in following sections.

2. RELATED WORK

The following sections summarize works done by others that are related to this project.

2.1 A Risk Sharing Proposal for Student Loans

The policy proposal titled "A Risk Sharing Proposal for Student Loans" by Tiffany Chou, Adam Looney and Tara Watson focuses on proposing a project "to introduce new and effective policy options" in order to improve economic opportunity for students in order to induce long-term prosperity. The proposal acknowledges the difficulty of students to repay their student loans at specific institutions. The paper basically proposes "an institutional accountability system" in order to align incentives of institutions with students and tax payers. The paper suggests various "risk-sharing" methods to analyze poor loan performance and overcome them through methods of mutual benefit in order to support both students and institutions that serve low-income students while also appropriately reimbursing the federal loan programs.

The following is a link to the article:

http://www.hamiltonproject.org/assets/files/risk_sharing_proposal_student_loans_pp.pdf

2.2 Is High Student Loan Debt Always a Problem?

This policy brief was written by Constantine Yannelis and Adam Looney in 2016 in conjunction with the Stanford Institute for Economic Policy Research. This short briefing summarizes key findings in the dataset offered. The primary finding being that students with high loan balances also tend to earn more. Therefore, the loan balance alone will not provide an accurate depiction of the student's financial health. Contrary to my initial hypothesis, students with low loan balances, less than $10,000, default on their loans 5 times as much in comparison to their high loan counterparts. This makes sense, as high loan borrowers typically went to very selective institutions and likely went on to acquire a graduate degree (i.e PhD, JD, MBA).

"Labor market outcomes, like unemployment or low earnings, provide a more direct measure of economic hardship" (Looney, A., & Yannelis, C. (2016)). This briefing suggests students should spend more time researching their field of study's opportunities for employment and potential income in that field. A $50,000 investment may be worth it if you have the potential to make $110,000 upon graduation.

The following is a link to the article:

http://siepr.stanford.edu/sites/default/files/publications/PolicyBrief-July16.pdf

2.3 A Crisis in Student Loans? How Changes in the Characteristics of Borrowers and in the Institutions They Attended Contributed to Rising Loan Defaults

This article, written by Adam Looney and Constantine Yannelis in 2015, "examines the rise in student loan delinquency and default [A Crisis in Student Loans…]" by analyzing data gathered from the U.S. Department of Education. This data was collected using earning records produced from tax records and describes federal student borrowing habits. The research within the article found that the increase of student loan default is a consequence of borrowers attending for-profit schools, non-selective schools, and community colleges. Of these institutions, for-profit and non-selective schools are primarily responsible for the increase in student loan default.

The following is a link to the article:

https://www.brookings.edu/wp-content/uploads/2015/09/LooneyTextFall15BPEA.pdf

3. DATA SET AND FEATURES

The data set to be used in this project is composed of a series of data tables containing higher education student financials data produced by the Federal Student Aid (FSA) which is an office of the U.S. Department of Education. The data includes populations of undergraduate, graduate and parent borrowers, as well as their respective loan borrowed

amounts and repayments. These are tabulated by institution of origin, meaning all the data including repayments rates are aggregated per institution. In addition to the aforementioned attributes, other example attributes of the data set include the following: ethnic class, percentage of completion, independent and dependent borrowers count, Adjusted Gross Income (AGI) for independent, dependent and parent borrowers, borrowers with Pell grant, median ages of borrowers, etc.

3.1 Tools

The following is a list of tools that will be used throughout the project. Tools or packages may be changed or added depending on project needs.

- Excel
- Anaconda
- Jupyter Notebook
- Python Programming Language and the following packages:
    - Numpy, Pandas, Matplotlib, Scipy, Sklearn, Orange

4. METHODS AND MODELS

4.1 Classification of Repayment Risk

The following data will be used to determine the risk of repayment for each institution.

- Group 1: Overall Outstanding Principal Balance
- Group 2: Overall Repayment Rate
- Group 3: % Increased Balance Borrowers
- Group 4: Defaulted Balance

Group 1 consists of the following data items, and describes the outstanding principal balance of each school as of September of the respective year. If an institution has a history of less risky lending habits, it is expected that these values should decrease over time.

- Overall Outstanding Principal Balance 5 YR Cohort (FY 2014)
- Overall Outstanding Principal Balance 4 YR Cohort (FY 2013)
- Overall Outstanding Principal Balance 3 YR Cohort (FY 2012)
- Overall Outstanding Principal Balance 2 YR Cohort (FY 2011)
- Overall Outstanding Principal Balance 1 YR Cohort (FY 2010)

Group 2 consists of the following data items, and describes the share of aggregate balance entering repayment repaid by cohort of the respective year. If an institution has a history of less risky lending habits, it is expected that this value should increase over time.

- Overall Repayment Rate 5 YR Cohort (FY 2014)
- Overall Repayment Rate 4 YR Cohort (FY 2013)
- Overall Repayment Rate 3 YR Cohort (FY 2012)
- Overall Repayment Rate 2 YR Cohort (FY 2011)
- Overall Repayment Rate 1 YR Cohort (FY 2010)

Group 3 consists of the following data items, and describes the share of borrowers whose current principal balance exceeds original principal balance. If an institution has a history of less risky lending habits, it is expected that these values should decrease over time.

- % Increased Balance Borrowers 2013-2014
- % Increased Balance Borrowers 2012-2013
- % Increased Balance Borrowers 2011-2012
- % Increased Balance Borrowers 2010-2011
- % Increased Balance Borrowers 2009-2010

Group 4 consists of the following data items, and describes the balance of loans currently in default. If an institution has a history of less risky lending habits, it is expected that these values should decrease over time.

- Defaulted Balance 2013-14

- Defaulted Balance 2012-13
- Defaulted Balance 2011-12
- Defaulted Balance 2010-11
- Defaulted Balance 2009-10

Linear regression analysis can be used to classify each school's risk level. If we find the line of best fit for each group, where X is the year and Y is the value for the group's school at the year, the slope of each line will determine potential risk according to Table 1.

Table 1: Slope Risk Assessment

ID	Risky	Not Risky
Group 1	Positive Slope	Negative Slope
Group 2	Negative Slope	Positive Slope
Group 3	Positive Slope	Negative Slope
Group 4	Positive Slope	Negative Slope

From here, schools can be classified as risky or not risky. Empirical analysis will be used to determine the best method of classification, as well as the different levels of risk that will be used for final classification.

4.2 Predictive Reasoning Capability Measurement

After the schools have been classified, the predictive reasoning capability of the attributes will need to be measured. Possibility of deriving the cause of a prediction will be determined by correlating each attribute of the data with risk classification. If a positive or negative correlation exists between the attribute and risk classification, then the data can be used to reason the prediction a student's probability of repaying a loan given certain circumstances – for this dataset, the circumstances are defined by the attributes of the school.

4.3 Predictive Model Creation

Finally, a predictive model can be created to predict a student's probability of repaying a loan given the attendance of a post-secondary school. This predictive model will exercise supervised learning, and will be done using a Decision Tree Classifier.

5. RESULTS AND DISCUSSION

The following sections detail the results of the each task defined in the Methods section.

5.1 Classification of Repayment Risk

5.1.1 *Data Preprocessing*

Before classification could be performed, the data set had to be cleaned. All values the were declared above or below the required threshold values of the attribute were set to 0 or 1, respectively.

Additionally, Group 1 (Overall Outstanding Principal Balance) and Group 4 (Defaulted Balance) were averaged to reduce the bias in the disparity in the borrow population size for each university.

5.1.2 *Borrower Risk Point Assignment Calculation*

The slope for each Group over time was then calculated in Excel. The quartiles for each Group are in the following tables.

Table 1: Average Outstanding Principal Balance Slope Quartiles (Group 1)

Quartile	Slope	Note
0	-1205.41875	min
1	156.26444	25th percentile
2	309.85581	50th percentile
3	737.790637	75th percentile
4	8771.5414	max

Table 2: Repayment Rate Slope Quartiles (Group 2)

Quartile	Slope	Note
0	-0.193532635	min
1	-0.057626055	25th percentile
2	-0.040448138	50th percentile
3	-0.022066621	75th percentile
4	0.071376503	max

Table 3: % Increased Balance Borrowers Slope Quartiles (Group 3)

Quartile	Slope	Note
0	-0.208695652	min
1	0.017511521	25th percentile
2	0.0405	50th percentile
3	0.063650307	75th percentile
4	0.188372093	max

Table 4: Defaulted Balance Slope Quartiles (Group 4)

Quartile	Slope	Note
0	-2812.146667	min
1	-311.7325522	25th percentile
2	-172.6142857	50th percentile
3	-88.06518096	75th percentile
4	929.468	max

Each quartile range was assigned a point value as show in Table 5.

Table 5: Quartile Point Assignment

Quartile	Point Value
0	0
1	0.25
2	0.5
3	0.75
4	1

The Borrower Risk Assignment for each school was then calculated with the following equation.

$$Borrower\ Risk\ = \frac{\sum_{k=0}^{n} Group\ Risk_n}{n}$$

5.1.3 *Borrower Risk Point Assignment Results*

The distribution of the Borrower Risk Assignment is shown in Figure 1.

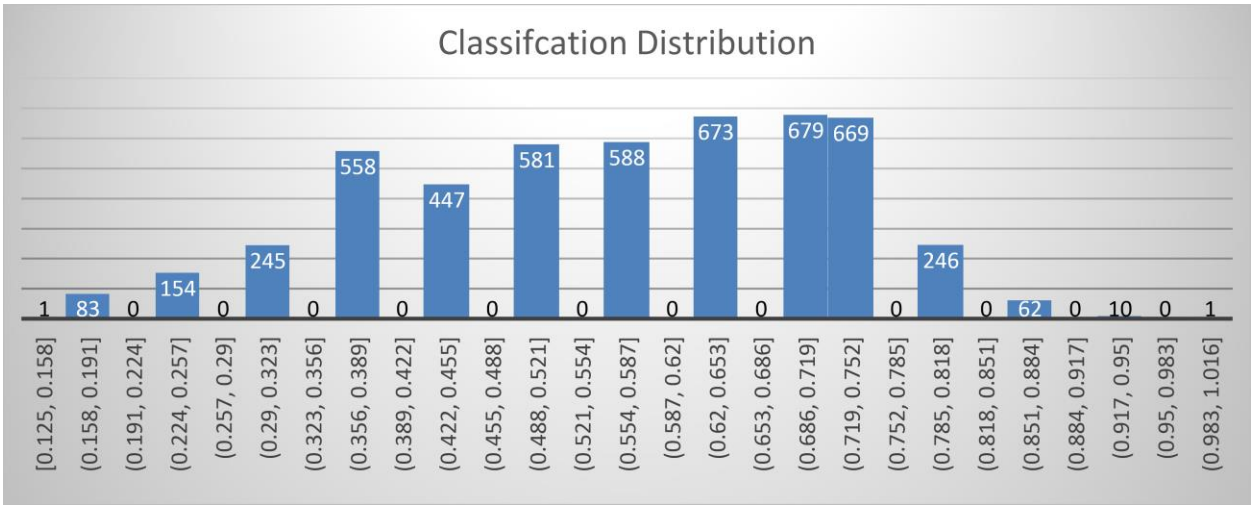

Figure 1: Borrower Risk Assignment Distribution

Quartiles of the Borrower Risk Assignment were used to assign the final classification of Borrower Risk for each school, as shown in Table 7.

Table 6: Borrower Classification

Quartile	Borrower Risk Assignment	Type	Count of Schools	Note
0	0.125	No Risk	1	min
1	0.438	Less Risk	1487	25th percentile
2	0.563	Moderate Risk	1169	50th percentile
3	0.688	More Risk	1352	75th percentile
4	1	Most Risk	988	max

5.1.4

5.2 Predictive Reasoning Capability Measurements

Using the repayment risk calculated above, the next step is to use linear regression to visualize the relationship between risk and the following attributes:

- % Completions Any School
- % Completions Same School
- % Independent Borrower Count
- % Dependent Borrower Count
- Median Independent Student AGI
- % Independent Borrowers with AGI < $30K
- Median Dependent Parent AGI
- % Dependent Borrowers with AGI < $30K
- % Borrowers with a Pell Grant
- Median Age of Dependent Borrowers at Maturity
- Median Age of Independent Borrowers at Maturity
- Mean Balance
- Median Balance

5.2.1 *Correlation Analysis*

After completing the classification of repayment risk, a correlation analysis was completed on the attributes listed above in the dataset. The attributes with moderate positive correlation were:

- Median Dependent Parent AGI

    - R-Value = 0.39
- % Borrowers without a Pell Grant
    - R-Value = 0.47

Figure 2 and Figure 3 are scatter plots with trendlines of the calculated risk vs Median Dependent Parent AGI and % Borrowers without a Pell Grant respectively.

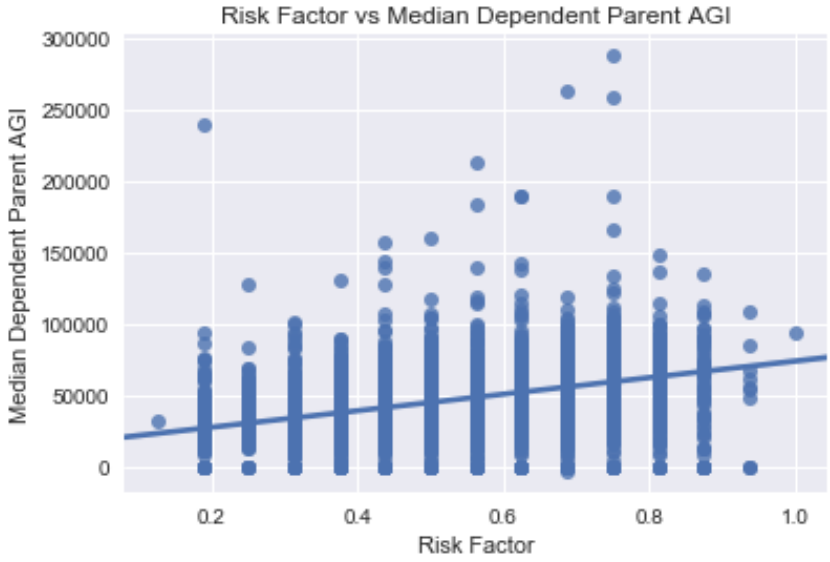

Figure 2: Risk Factor vs Median Dependent Parent AGI

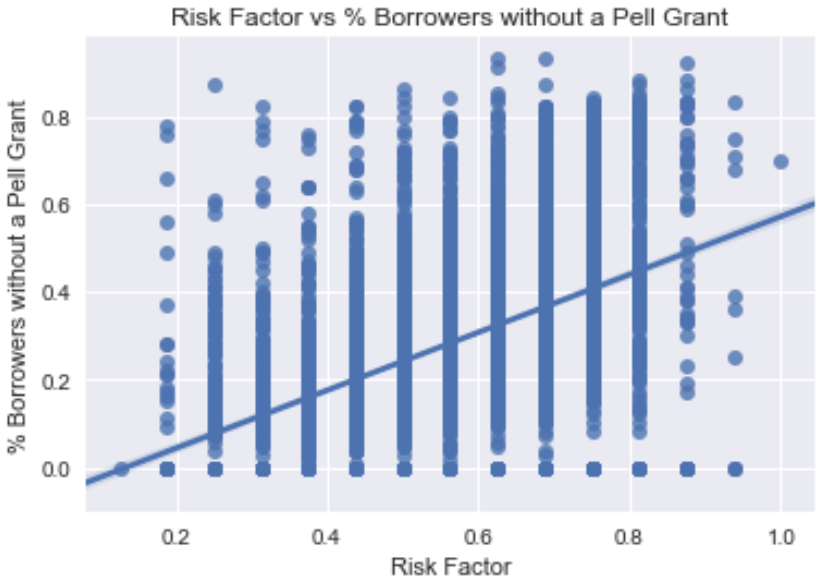

Figure 3: Risk Factor vs %Borrowers without a Pell Grant

Attributes with a very weak positive correlation were:

- % Dependent Borrower Count
    - R-Value = 0.21
- Mean Balance
    - R-Value = 0.22

Figure 4 and Figure 5 are scatter plots with trendlines of the calculated risk vs % Dependent Borrower Count and Mean Balance respectively.

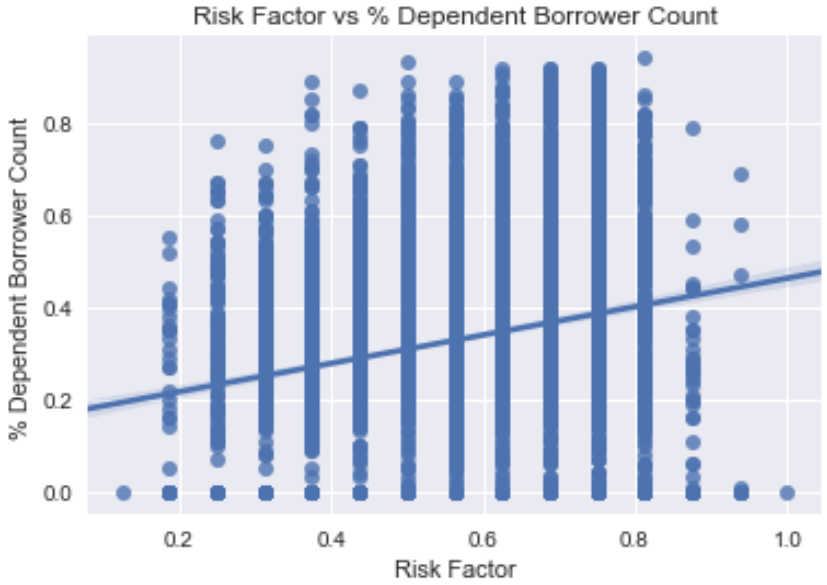

Figure 4: Risk Factor vs % Dependent Borrower Count

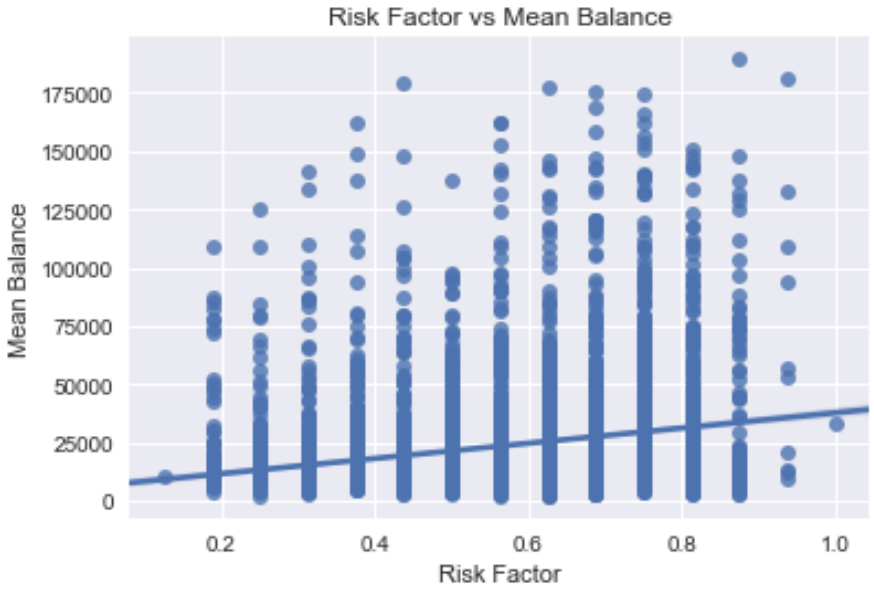

Figure 5: Risk Factor vs Mean Balance

Attributes showing a very weak negative correlation include:

- % Dependent Borrowers with AGI < $30k
    - R-Value = -0.25

Figure 6 is a scatter plot with a trendline of the calculated risk vs % Dependent Borrowers with AGI < $30k.

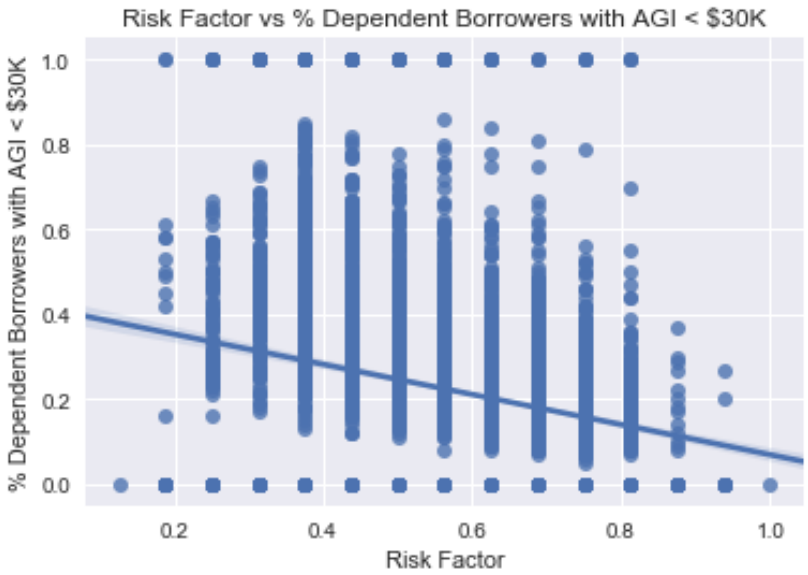

Figure 6: Risk Factor vs % Dependent Borrowers with AGI < $30k

5.2.2 *Association Analysis*

In addition to correlation analysis, it was decided to perform association analysis to determine the correlation between the risk classification and the other non-numerical attributes in the data set. The non-numerical values describing the institutions in the data set include the following: Eligibility for financial aid, certification, type and control and the newly calculated borrowing risk classification.

Therefore, the Orange python package was used to perform this analysis. The rules were generated from the data using minimum support as 0.1 and minimum confidence as 0.2. Using Orange's capabilities, the lift measure was also computed for each of rules generated. The lift was the predominant measure in this portion of the project since it is the measure of dependent/correlated events. The association rules generated were also sorted for clear visualization. The following is the resulting table of association rules.

Table 6: Association Rules for Borrowing Risk

supp	conf	lift	rule
0.164	0.881	1.233	Borrowing Risk=Most Risk ELIGIND=Y -> CERTCD=C

0.164	0.23	1.233	CERTCD=C -> Borrowing Risk=Most Risk ELIGIND=Y
0.215	0.852	1.192	Borrowing Risk=More Risk ELIGIND=Y -> CERTCD=C
0.215	0.301	1.192	CERTCD=C -> Borrowing Risk=More Risk ELIGIND=Y
0.164	0.832	1.165	Borrowing Risk=Most Risk -> CERTCD=C
0.164	0.832	1.165	Borrowing Risk=Most Risk -> ELIGIND=Y CERTCD=C
0.164	0.23	1.165	CERTCD=C -> Borrowing Risk=Most Risk
0.164	0.23	1.165	ELIGIND=Y CERTCD=C -> Borrowing Risk=Most Risk
0.215	1	1.142	Borrowing Risk=More Risk CERTCD=C -> ELIGIND=Y
0.215	0.245	1.142	ELIGIND=Y -> Borrowing Risk=More Risk CERTCD=C
0.173	1	1.142	Borrowing Risk=Less Risk CERTCD=C -> ELIGIND=Y
0.164	1	1.142	Borrowing Risk=Most Risk CERTCD=C -> ELIGIND=Y
0.162	1	1.142	Borrowing Risk=Moderate Risk CERTCD=C -> ELIGIND=Y
0.162	0.795	1.113	Borrowing Risk=Moderate Risk ELIGIND=Y -> CERTCD=C
0.162	0.226	1.113	CERTCD=C -> Borrowing Risk=Moderate Risk ELIGIND=Y
0.215	0.794	1.112	Borrowing Risk=More Risk -> ELIGIND=Y CERTCD=C
0.215	0.794	1.112	Borrowing Risk=More Risk -> CERTCD=C
0.215	0.301	1.112	ELIGIND=Y CERTCD=C -> Borrowing Risk=More Risk
0.215	0.301	1.112	CERTCD=C -> Borrowing Risk=More Risk
0.187	0.944	1.078	Borrowing Risk=Most Risk -> ELIGIND=Y
0.187	0.213	1.078	ELIGIND=Y -> Borrowing Risk=Most Risk
0.252	0.933	1.065	Borrowing Risk=More Risk -> ELIGIND=Y
0.252	0.288	1.065	ELIGIND=Y -> Borrowing Risk=More Risk
0.173	0.742	1.039	Borrowing Risk=Less Risk ELIGIND=Y -> CERTCD=C
0.173	0.242	1.039	CERTCD=C -> Borrowing Risk=Less Risk ELIGIND=Y

0.203	0.869	0.992	Borrowing Risk=Moderate Risk -> ELIGIND=Y
0.203	0.232	0.992	ELIGIND=Y -> Borrowing Risk=Moderate Risk
0.162	0.691	0.967	Borrowing Risk=Moderate Risk -> CERTCD=C
0.162	0.691	0.967	Borrowing Risk=Moderate Risk -> ELIGIND=Y CERTCD=C
0.162	0.226	0.967	CERTCD=C -> Borrowing Risk=Moderate Risk
0.162	0.226	0.967	ELIGIND=Y CERTCD=C -> Borrowing Risk=Moderate Risk
0.233	0.783	0.895	Borrowing Risk=Less Risk -> ELIGIND=Y
0.233	0.266	0.895	ELIGIND=Y -> Borrowing Risk=Less Risk
0.173	0.582	0.814	Borrowing Risk=Less Risk -> CERTCD=C
0.173	0.582	0.814	Borrowing Risk=Less Risk -> ELIGIND=Y CERTCD=C
0.173	0.242	0.814	CERTCD=C -> Borrowing Risk=Less Risk
0.173	0.242	0.814	ELIGIND=Y CERTCD=C -> Borrowing Risk=Less Risk

Based on the initial conditions, it was determined that the type and control attribute was found to not be a strong association rule. In other words, in this initial assessment for the project, type and control will not be a deciding attribute in our prediction model. Therefore, the rest of the attributes, including eligibility for aid and certification, were found to be part of the strong association rules with lifts that display either negative and positive correlation, depending on the association rule. The next step in this portion of the project will be to carefully analyze each of the association rules to see what further details can be extracted.

5.3 Predictive Model Creation

The predictive model has not yet created. The initial steps for the creation are being examined for further refinement.

6. CONCLUSION

The conclusion will be completed as soon as the overall project reaches a considerable level of maturity and completion.

APPENDIX A – TEAM CONTRIBUTIONS

Table 7: Team Contributions

Team Member	Contribution
Misti Stevens	Pre-processing and Classification
Omar Waller	Correlation Analysis
Alfredo Antolinez	Association Rule Generation and Analysis