Addition Financial Credit Limit Analytics Competition

Hayden D. Hampton
Department of Statistics and Data Science, University of Central Florida

***Abstract*—In credit lending, a common practice for management is to identify members who qualify for a credit line increase and determine an appropriate amount. For this analysis, we will consider data on the credit card portfolio at Addition Financial Credit Union. The goal is to review the data on a per member basis, and develop a predictive model that determines if a member should receive a credit limit adjustment and by what amount.**

***Index Terms*—predictive analytics, portfolio optimization, credit lending, banking, financial services.**

I. Executive Summary

A. Introduction

In credit lending, a common practice for management is to identify members who qualify for a credit line increase and determine an appropriate amount. For this analysis, we will consider data on the credit card portfolio at Addition Financial Credit Union. The goal is to review the data on a per member basis, and develop a predictive model that determines if a member should receive a credit limit adjustment and by what amount.

B. Modeling Approach

The modeling approach undertaken to solve this problem initially involved a two-stage feature engineering phase. The first set of features were developed using the four distinct datasets provided, and the second set of features were developed after a centralized database was created. These features include borrower risk profiles, duration of credit line, credit line usage patterns, member payment patterns, and other behavioral patterns that often correlate with undesirable risk characteristics. Features related to external debt and whether members lacked prior credit history were also considered.

We approached the problem of predicting if a member should receive a credit limit adjustment and by what amount by developing a custom model that was both intuitive and interpretable by design. By focusing on extensive feature engineering and using precise imputation techniques, we designed a tree-based model which tracks risk characteristics on a per member basis that were deemed significant in credit worthiness evaluation. By updating a numeric quantity based on a member's particular risk characteristics which we call the risk score, we were able to predict whether a given member is qualified for a credit limit increase and by what amount.

C. Results

Model tuning and validation was approached by incorporating the assumptions that were made when initially developing the model. It was assumed that summarizing relevant customer information and reducing it into a usable form would result in the ability to intuitively segment customers by risk characteristics and determine which members are qualified for a credit line increase. Based on the model's underlying assumptions, we weighted the risk characteristics that we assumed to be significant in model development. Then we tuned three additional model parameters that impact the populations of interest while working in conjunction with the risk score. Finally we analyzed which members were being approved, as well as those being rejected, and were able to validate our model was working in an intuitive manner after adequate tuning. The model is denying credit line increases to clients with high balances that have overall undesirable risk characteristics. Furthermore, it is approving credit lines increases on a per-member basis with an appropriate amount for those clients with overall desirable risk characteristics.

D. Recommendations

Due to the nature of the data, as time passes and various aspects of the portfolio change the model will require further adjustment. If economic conditions such as inflation continue to rise, and the current pandemic continues to hinder economic growth, management should review the model's parameters and continue tuning them to obtain optimal performance. Based on the company's risk appetite and desire to grow the portfolio, the mechanism by which the risk score is optimized can be adjusted to reach the desired outcome. Using historical data on prior credit line increases, or by using this model and tracking clients over time, the company has the ability to meet its desired goals.

II. Data Analysis

A. Introduction

In credit lending, a common practice for management is to identify members who qualify for a credit line increase and determine an appropriate amount. Financial institutions are now carefully tuning their lending strategies for evaluating borrower risk in an effort to predict which members are most likely to pay back debt in a timely manner with the intent to improve customer retention and increase profitability for the company. This enables the managers to selectively offer credit line increases to those members with desirable risk characteristics over others with undesirable risk characteristics. For this analysis, we will consider data on the credit card portfolio at Addition Financial Credit Union. The goal is to review the data on a per member basis, and develop a predictive model that predicts if a member should receive a credit limit adjustment and by what amount.

For this analysis, we will consider the four datasets provided by Addition Financial Credit Union. These datasets provide a variety of information on the credit card portfolio at Addition

Financial. The goal of this analysis is to review the data on a per member basis, and develop a predictive model that determines if a member should receive a credit limit adjustment and by what amount. This enables qualified members to have greater access to credit which drives retention and customer loyalty. Through customer usage, it also increases the financial institution's revenue due to Visa © interchange income and interest income for the company.

The four original datasets were comprised of information on a per member basis related to aggregated unsecured debt, company requests of credit bureau reports, credit card rate changes, and member financial information. These four datasets were linked by a unique customer identification number, which we will refer to as the client ID, and this was ultimately used to create a centralized database containing all pertinent information that was used in the analysis.

The dataset with the members' aggregated unsecured debt information consisted of 6 features with a total of 38,072 observations. Aside from the client ID, these were all quantitative features. The dataset related to the frequency of credit card rate changes consisted of 3 features with a total of 199,116 observations. Aside from the client ID, there was one numeric feature representing the credit card rate and one date feature corresponding to the time at which each respective rate change occurred. The dataset related to the number of times the financial institution pulled the client's credit report from a rating agency consisted of 3 features with a total of 249,070 observations. Aside from the client ID, there was one numeric feature representing the credit bureau score and one date feature corresponding to the time the inquiry was made. The dataset related to the members' financial information from the company consisted of 43 features with a total of 351,835 observations. Aside from the client ID, there were 30 quantitative features and 12 date features.

B. Data Description

An initial challenge with the provided data is that some members held multiple unsecured credit cards with the financial institution. To circumvent this complication, this analysis only focused on members with a single unsecured credit card. In total there were 640 members identified as holding multiple credit lines, where the majority of these members had 2 lines of credit and 9 members had 3 lines of credit. One member was identified as having 8 lines of credit. After creating a centralized database using all four datasets, we ultimately found that there were 28,028 members with a single unsecured credit card that we were able to perform prediction on.

C. Data Preparation

In the dataset containing the members' aggregated unsecured debt information, the company instructed us to treat any missing observations as zero. Therefore, if any of the features including the number of loans, the total loan debt, the number of credit cards, the total credit limit, or the total unsecured debt contained missing values, they were replaced with a zero value. It was also discovered that there were many duplicate records when filtering by the client ID. The company suggested that these duplicates may be an indication of multiple accounts for a given member, but this did not seem to be in agreement with the information we extracted from the other datasets which provided a less ambiguous method of identifying members with multiple lines of credit. The majority of the redundant information in this dataset was populated in the database lacking any useful information and largely contained missing values. Therefore, it was dealt with accordingly and any redundant information was removed.

In the dataset related to the members' financial information, there were features related to predictive hardship, predictive attrition, and predictive growth which the company instructed us to remove due to data quality issues. These features along with their corresponding dates were removed from the dataset. Additionally, the company instructed us that there were data quality issues with the primary cardholder personal monthly income and the primary cardholder disposable monthly income features, and these were also removed from the dataset. A group of 276 members were identified as having credit scores of 1 which the company informed us represented members with secure credit cards. Since this analysis is only focused on members with unsecured credit cards, these members were removed from the dataset. Another group of 112 members identified as having a single unsecured credit card were flagged due to conflicting data on various levels in the original data set, and were partitioned into a separate data set. This subset of members will require individualized attention by the company to identify the source of the ambiguous data.

The dataset related to the frequency by which the company requested a credit bureau report had no missing data. Similarly, the data set related to the frequency of rate adjustment for the member's credit card had no missing data.

D. Feature Engineering

Before the four datasets were combined into a centralized database, some feature engineering was required within the individual data sets. This was the first stage of feature engineering done prior to modeling, but we ultimately took a two-fold approach to feature engineering in this analysis. After all the desired features were created in the four distinct data sets, the second stage involved developing features in the centralized database.

In the dataset related to the number of times the member's credit report was pulled, several features were developed to extract the information of interest. On a per member basis, we were interested in determining the initial credit score and the date corresponding to when the report was generated, the frequency count related to how many credit bureau reports were pulled, the average of the credit scores excluding the first, and the latest credit score and the date corresponding to when the report was generated. Then we developed a feature which indicated whether there was movement in the credit score on a per member basis from the initial credit score compared with the average of all subsequent credit scores. This binary flag was later used to quantify the movement that occurred. If

only a single score was available for a given member, it was indicated appropriately by the flag.

In the dataset related to the number of times the credit card rate changed, we developed similar features. On a per member basis, we were interested in the initial rate, the frequency count corresponding to how many rate changes occurred, the average of the rates excluding the first, and the latest rate. Then we developed a feature which indicated whether there was movement in the rate on a per member basis from the initial rate compared with the average of all subsequent rates. This binary flag was later used to quantify the movement that occurred. If only a single rate was available for a given member, it was indicated appropriately by the flag.

In the dataset related to the members' financial information, we initially created a feature to represent the number of unsecured credit cards each member had. After a careful examination of the database, a pattern was identified that enabled unique credit line identification. First, a filter was applied to eliminate any accounts with a lifetime high balance less than or equal to 0. Then we sorted all observations by the client ID, the date when the credit card was opened, and the date when the credit card expires. Then we flagged any duplicates of these three features which enabled us to identify all of the unique accounts each member had with the company. Then we partitioned members with multiple lines of credit and removed any duplicate observations of members with a single line of credit that may have been populated into the database inadvertently.

With the first stage of the feature engineering complete, we combined all four data sets using the client ID. However, there were members represented in one or more of these four data sets but not present in the others which resulted in a large quantity of missing values. For example, if a member had their credit line amount listed in the unsecured debt data set, but it was missing in the members' financial data set, it was updated in the appropriate missing field.

With all the data in one centralized database, we then performed the second stage of feature engineering. This involved developing features related to the members' behavior in a variety of ways. Due to the structure of the data, there was a loss of granularity because we did not have multiple account updates over time on a per member basis. Therefore, it was important to develop features that shed light on the members' behavioral patterns.

The first behavioral feature developed corresponds to the members' credit line usage and repayment patterns. We were interested in determining if a member is a convenience user, in other words the member pays off their outstanding balance at the end of every month, or if a member is a revolving user, in other words the member maintains an outstanding balance and pays any accumulated interest over time. To create this feature, we compared the prior year's amount of reportable interest against the year to date purchase interest amount on a per member basis. The best we were able to do in this regard was to say that for accounts approximately two years and older, if these two interest quantities are both zero then we have a convenience user based on the available data which covers approximately two concurrent years. Additionally, this feature also assumes there was activity on the card during that time.

The second behavioral feature developed corresponds to whether there was cash advance activity on a per member basis. Cash advances in conjunction with other behavioral patterns can be a sign of undesirable risk characteristics. Due to the way the database generates cash advance activity, this was not a straight forward feature to develop. In fact, it required three separate approaches to develop this one feature properly. We initially created an indicator flag that activates when a given member participates in cash advance activity. First, if a member had performed a cash advance within the current year, then it should be represented under the year to date cash advance amount feature and the year to date cash advance count feature, respectively. Second, for older cash advances it should be represented under the last cash advance date feature. However, it was identified that certain cash advances were not being accounted for correctly when only the two aforementioned techniques were used. A clever workaround we used to identify the cash advances these two approaches would have otherwise missed was to check if a member's lifetime high balance exceeded their credit line amount. With this approach, we are assuming that there was not a prior decrease in the member's credit line. We discovered that the database populates itself in such a way that this is a third way to identify older cash advance activity not otherwise accounted for. Since cash advance activity can be a strong behavioral characteristic of interest when segmenting customers, it was important we accounted for all activity from the data we were given access to.

The third behavioral feature corresponds to the member's payment patterns. We developed a feature that represents the credit utilization ratio on a per member basis. This was calculated by dividing the current account balance by the maximum amount of credit available. Then we created a flag which indicated on a per member basis if the credit utilization ratio was higher than thirty percent, a threshold which is a common industry standard. If this ratio was unknown, this was also flagged appropriately.

Then we created a feature that corresponds to whether a member has an unsecured loan in addition to the unsecured credit card line. This is later used in the modeling as an indicator of additional debt the members' have which may impact their repayment ability if they are offered a credit line increase. We did not include cash advances or external loans into the total available debt because we were often unable to determine how much outstanding debt was owed given the current data.

From the raw data, there were multiple features related to delinquency that were of interest. We placed more emphasis on the number of times the cardholder's account cycled in a two cycle delinquent status since the open date (greater than 60 days) as opposed to the number of times the cardholder's account cycled in a one cycle delinquent status since the open date (between 30-59 days), but both were significant features

in our model. Instead of using the lifetime frequency count that was provided in the raw data, we created ratios that incorporated the age of the credit line and averaged the number of delinquencies over the lifetime of the account. We believed that analyzing the counts in their raw form could be misleading if the age of the account isn't factored in.

Due to the heterogeneous nature of this data, we decided to take a tiered pricing approach using customer segmentation. The first step in this process was identifying the most recent credit score on a per member basis. We assumed that all credit scores in the data were FICO credit scores, and did not differentiate between internal company credit scores and the FICO scores generated from the credit bureau reports. Then we created a borrower risk classification corresponding to the members' latest credit score. We used the standard five FICO credit score ranges where less than 580 is a poor credit score, 580 to 699 is a fair credit score, 670 to 739 is a good credit score, 740 to 799 is a very good credit score, and 800 to 850 is an exceptional credit score.

One issue that arose after creating the centralized database was the identification of members with no known credit score. These "credit-invisibles" are members who may have no credit history for a variety of reasons, and the cause was not possible to determine given the data provided. The lack of credit history on these members does not imply they have bad credit. In fact, they may be credit worthy but it is more difficult to determine this than for the other members with credit history. For this reason, these members were isolated from the other members, and further analysis will need to be conducted by the financial institution to determine if these members should be granted a credit line increase. The data that was available on these members often left too many important characteristics unknown to adequately determine if a credit line increase was warranted. Loans to credit invisible members should be made with caution, and will require more information than was provided in this data.

E. Exploratory Analysis

To obtain a broad view of the data provided, different visualizations were created that showed some interesting aspects of the data. First we wanted to ensure that we grouped the members based on credit line amount into appropriate buckets to be used in our modeling strategy. We can see in Figure 1 that the buckets we chose all have approximately the same number of members.

We were also interested in performing a vintage analysis in an effort to understand the repayment trends for accounts of different ages. By analyzing the portfolio from the perspective of the age of the accounts, we can see how the lifetime delinquency count for greater than 60 days varied. We observed that there was a large spike in delinquency for accounts that were between approximately 6 and 9 years old as shown in Figure 2. This portfolio trend was an unexpected result, and was one reason a heavy penalty was placed on accounts showcasing significant delinquency behavior.

III. Modeling

A. Modeling Approach

Since the centralized database was comprised of multiple datasets, the amount of missing data was a key reason we did not approach this problem using conventional supervised or unsupervised learning techniques. It was stressed by the company that an intuitive and interpretable model was desired, which seemed to rule out unsupervised learning. For example, if we took an unsupervised learning approach to identify clusters it would be challenging to justify why clients were being clustered into their respective groups. Another concern was that imputation on certain features did not seem appropriate given how much weight some features are traditionally given when evaluating credit worthiness. Common imputation techniques attempted on this data often produced results that were difficult to justify and lacked interpretability. Additionally, the provided data did not have any meaningful label that could be used for training to predict credit line increases in a supervised learning setting. If a label was to be developed with the current data, a large concern was that the model would generalize poorly since unseen data may have a large amount of missing data, as we encountered in the supplied data, and the complications related to imputation would persist.

For these reasons, we developed a custom model that approached this problem in the most intuitive manner possible. We focused on extensive feature development and precise imputation techniques which ultimately enabled us to build a tree-based model. This model tracks risk characteristics on a per member basis that were deemed relevant to credit worthiness. Similar to the traditional scorecard modeling approach, we aimed to reduce the relevant information about customers into an intuitive outcome, but with the exception that in our model we used a numeric value, which we call the risk score, that goes through a tuning procedure as opposed to the more traditional ordered categories (or scores). When a risk score is calculated for each member, the borrower risk profile, amount of time the line of credit has been in use, and the credit line amount are considered before predicting whether a member should have a credit limit adjustment and by what amount. The maximum possible increase for a given member is also impacted by the duration of the credit line. If a member has been with the company for at least 5 years, we offered qualified members a larger increase than qualified members who have had an open credit line with the company for less than 5 years. The model also incorporates cutoff ratios and weighting parameters for additional model optimization. This approach enables management to fine tune the model over time by adjusting the parameters to meet company goals. Based on the company's risk appetite and whether they are interested in growing the portfolio or minimizing their credit exposure, the mechanism by which the risk score is calculated can be tuned for optimal performance. Incorporating historical data on prior credit line increases, or by using this model and tracking members over time, the company has the ability to meets its desired goals.

B. Statistical Modeling

On a high level, we can visualize this model using a tree diagram as shown in Figure 3, although this illustration heavily simplifies the mechanism by which the true model makes predictions. To accurately showcase the full decision tree this model uses would be difficult to represent on a single page. This simplified diagram also omits numerous features that penalize the risk score due to missing data. As previously mentioned, we did not believe performing imputation on a feature related to delinquency was appropriate, so we instead took the approach of creating a new feature which flags missing data in this field. The model checks if this flag was activated, and there is a penalization that occurs due to its presence. Similarly, there is a flag activated if the credit utilization ratio is unknown. If the data prevented us from determining if a member was a convenience user or credit card revolver, we chose to err on the side of caution and label the user as a credit card revolver.

The duration of credit line was set as an important feature in our model and it heavily influences the model behavior. There was some missing data for this feature, and to perform imputation we analyzed the known ages of accounts with respect to the lifetime frequency credit bureau reports pulled as well as changes in the credit card rate. Based on the borrower risk profile, different thresholds were used to determine if the missing age would be set to 5 years or 2.5 years, respectively. There was a trend that clients with low credit scores had high frequency counts in these two areas when the account age was less than 5 years old, so we thought it was prudent to raise this cutoff value for higher risk clients.

Additionally, members designated into the poor category with credit scores less than 580 were classified as sub-prime and were not eligible for a credit line increase at this time. We also considered any clients with accounts less than a year old as ineligible for a credit line increase due to lack of available historical data.

C. Model Comparison

Since we did not have access to data related to prior credit line increases, we had to consider model validation from a perspective other than traditional classification with known labels for comparison. We tuned and ultimately validated our model by using the assumptions we made when developing the model. We assumed that summarizing relevant customer information and reducing it into a usable form would result in the ability to intuitively segment customers by risk characteristics, and determine which members are qualified for a credit line increase. We had a few specific characteristics that we considered most relevant for this process. These included the borrow risk profile, which was dictated by the latest credit score, delinquency behavior, payment behavior, change in credit scores over time, change in credit card rates over time, and external debt. By analyzing which clients were being approved for credit line increases, as well as those that were being rejected, in a broad sense we were able to find suitable model parameters for the data and validate our model.

IV. Results and Conclusion

Model tuning and validation was approached by using the assumptions that were made when initially developing the model. It was assumed that summarizing relevant customer information and reducing it into a usable form would result in the ability to intuitively segment customers by risk characteristics and determine which members are qualified for a credit line increase. Based on the model's underlying assumptions, we weighted the risk characteristics that we assumed to be significant in model development. Then we tuned three additional model parameters that impact the populations of interest while working in conjunction with the risk score. Finally we analyzed which members were being approved, as well as those being rejected, and were able to validate our model was working in an intuitive manner after adequate tuning. The model is denying credit line increases to clients with high balances that appear to have overall undesirable risk characteristics. Furthermore, it is approving credit lines increases on a per-member basis with an appropriate amount for those clients with overall desirable risk characteristics.

Due to the nature of the data, as time passes and various aspects of the portfolio change the model will require further adjustment. If economic conditions such as inflation continue to rise, and the current pandemic continues to hinder economic growth, management should review the model's parameters and continue tuning them to obtain optimal performance. Based on the company's risk appetite and desire to grow the portfolio, the mechanism by which the risk score is optimized can be adjusted to reach the desired outcome. Using historical data on prior credit line increases or by using this model and tracking clients over time, the company has the ability to meet its desired goals.

V. Future Considerations

In the future it would be worthwhile to further develop the model so that the tuning can be done using a grid search and optimal portfolio performance be recorded as different parameter values are tested. The chosen parameters at this time were found to work best through trial and error.

Further analysis related to the credit invisible members is also warranted. The modeling approach outlined in this analysis may be suitable for these clients, but with some adjustments due to lack of credit history.

Figure 1. Frequency Distribution of Clients by Segmented Credit Line Amounts

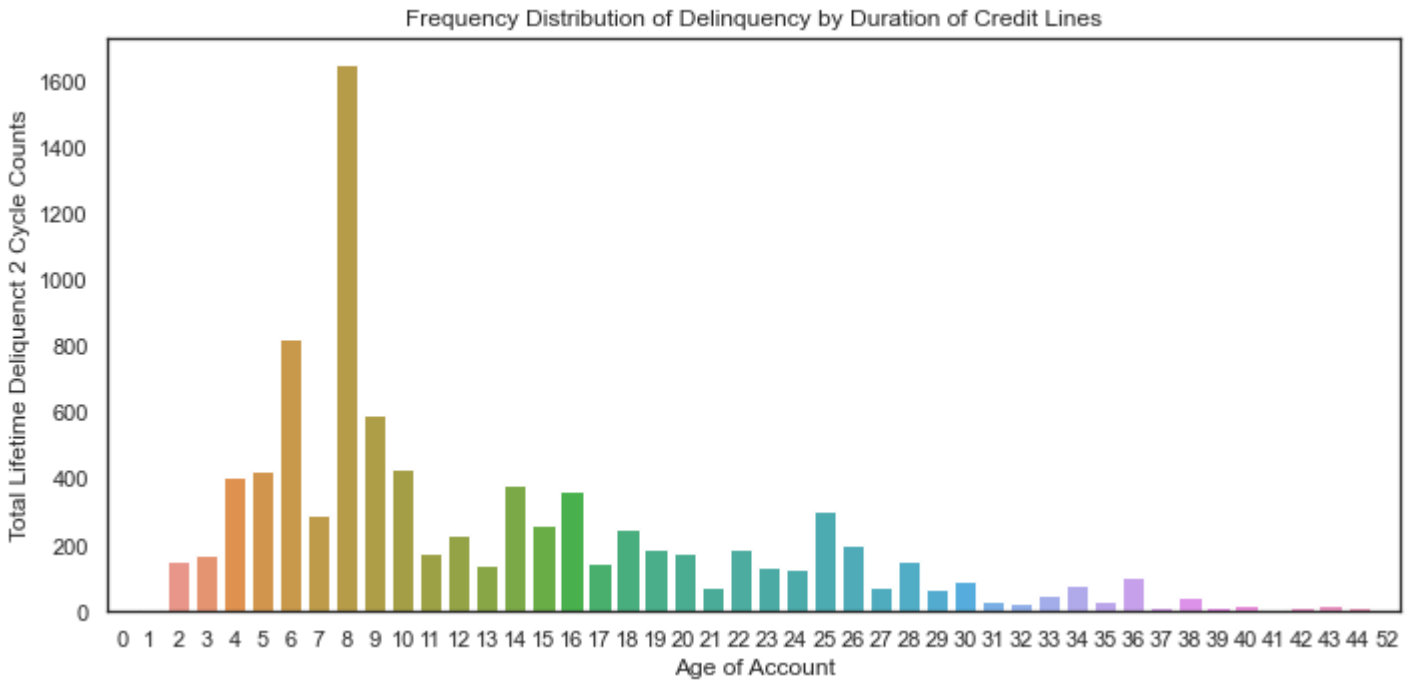

Figure 2. Vintage Analysis

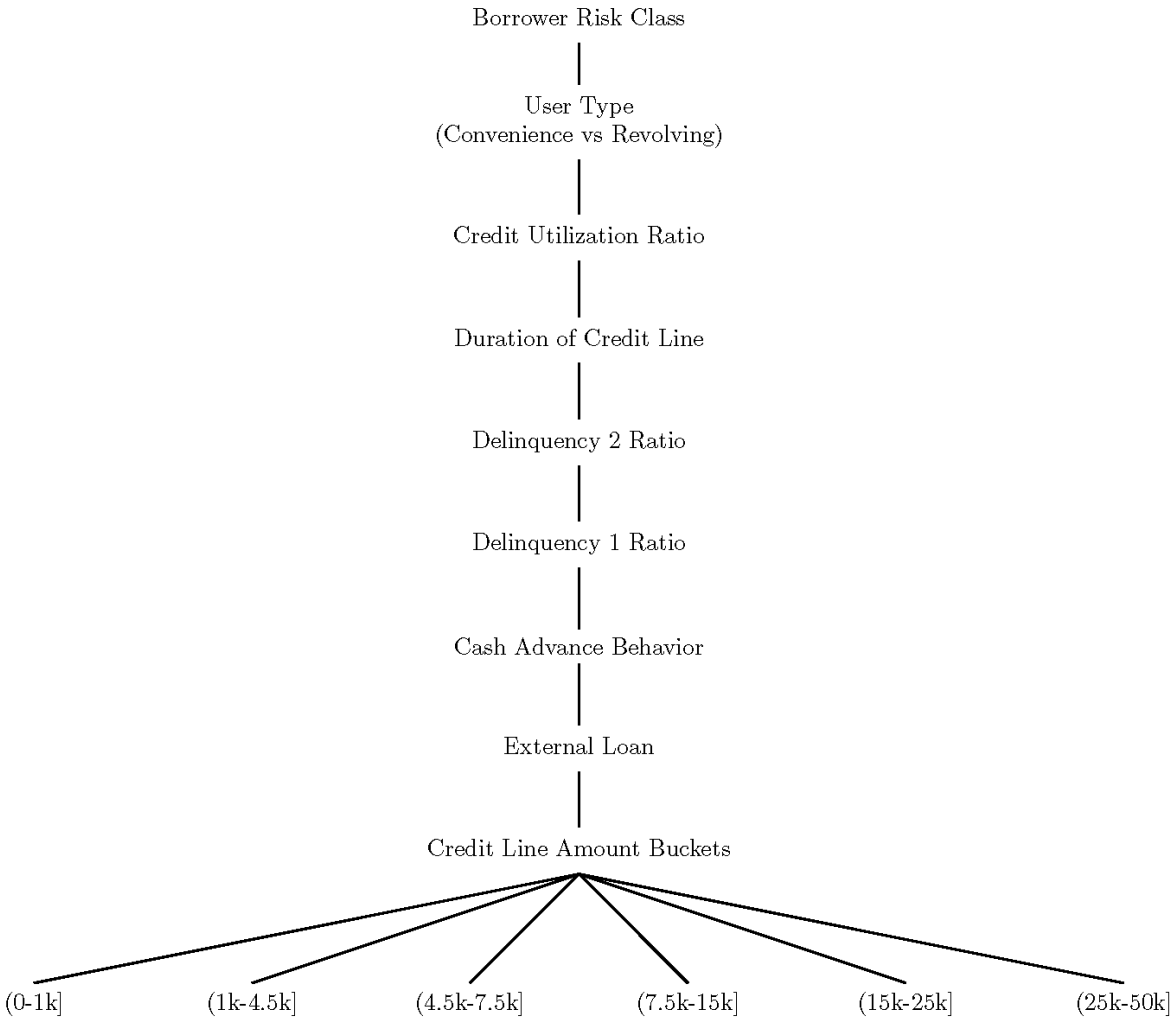

Figure 3. Predictive Model Diagram