Loan Default Detection through Customer Segmentation

Arabind Swain, Fiona Wu and Richa Sharma[1]

[1] *Emory University, CS534: Machine Learning*

(Dated: 4th December 2020)

Customer Segmentation is studied using K-means and K-mode clustering techniques, while feature relevance is calculated based on predictive models of Logistic and CART. Selection of the dataset, and problem statement of the Market Segmentation was done by Fiona, she also performed data visualization and Feature engineering. Correlation Graphs were implemented using Pearson method by Richa, who also developed the predictive models of Logistic Regression and CART and analyzed the most relevant features in them. Lastly, Arabind took the feedback from the predictive models and implemented the K-means and K-mode Clustering techniques.

I. INTRODUCTION

Customer Segmentation is about dividing customers into different groups based on their characteristics. It is extremely beneficial for companies to obtain this knowledge about their customers based on collected historical data. In this project, we use a loan data-set which contains a track record of individuals loan payment history and their characteristics such as age, income type, gender, education, etc. Customer segmentation is crucial for banks to decide whether or not they would issue a loan to a new or existing customer, and at which interest rates they are issuing the loan. It is also helpful to identify the best communication channels for different groups, therefore enhancing customer relationships and developing more personalized customer service. Most importantly, it helps banks to allocate resources and focus on the most profitable customer groups.

We first learn about the data-set through feature engineering, and feature selections. The target variable which links to prediction is the historical customer loan status, ranging from not having loans at all, paid off during that month, within 90 days overdue, to over 90 days overdue. Each customer has up to 5 years of loan payment track record. We then implement supervised learning such as logistic regression and CART, and unsupervised learning techniques such as K-Means and K-Modes to categorize customers into easy to default or unlikely to default baskets. Finally, we evaluate our model prediction accuracy and conclude.

II. KAGGLE DATASET AND FEATURE SELECTION

A. Dataset

The loan dataset comes from Kaggle. The target variable is loan status. We categorize loans which are within 90 days overdue as unlikely to default, otherwise there is a high chance of default. There are 17 explanatory features in the dataset such as gender, if the customer owns a car or house, their living conditions, annual income, education level, etc. Below are some feature visualizations. 67% of the population are women. Customers age ranges from 20 to 69 with a median of 43 years old. The annual household income distribution is highly skewed with a median of 160K. 51% of the customers obtain their income through working. Less than 30% of the customers pursue higher education or have completed their academic degree. 73% of the customers are in a marriage. Finally, 97% of the records belong to unlikely to default in this dataset.

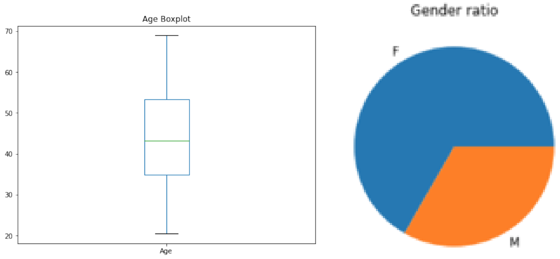


FIG. 1. Database visualization in terms of box plot and pie chart.

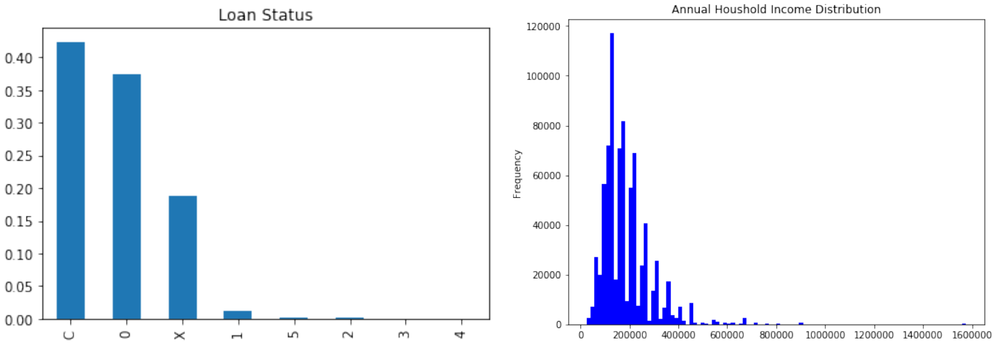
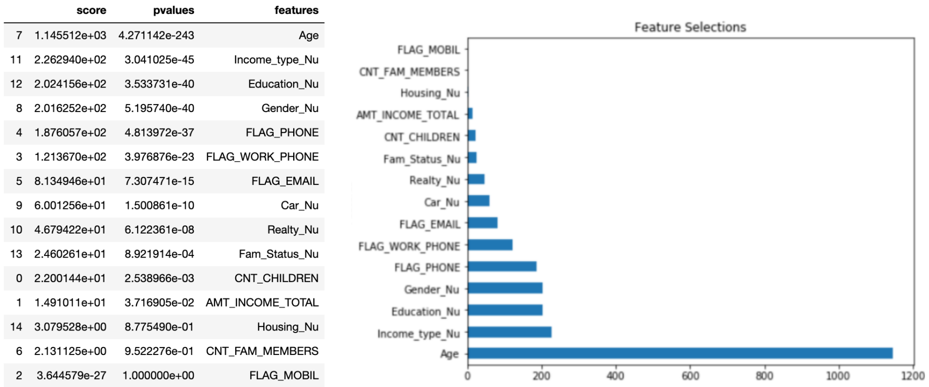


FIG. 2. Histogram for loan status and annual household income.



	score	pvalues	features
7	1.145512e+03	4.271142e-243	Age
11	2.262940e+02	3.041025e-45	Income_type_Nu
12	2.024156e+02	3.533731e-40	Education_Nu
8	2.016252e+02	5.195740e-40	Gender_Nu
4	1.876057e+02	4.813972e-37	FLAG_PHONE
3	1.213670e+02	3.976876e-23	FLAG_WORK_PHONE
5	8.134946e+01	7.307471e-15	FLAG_EMAIL
9	6.001256e+01	1.500861e-10	Car_Nu
10	4.679422e+01	6.122361e-08	Realty_Nu
13	2.460261e+01	8.921914e-04	Fam_Status_Nu
0	2.200144e+01	2.538966e-03	CNT_CHILDREN
1	1.491011e+01	3.716905e-02	AMT_INCOME_TOTAL
14	3.079528e+00	8.775490e-01	Housing_Nu
6	2.131125e+00	9.522276e-01	CNT_FAM_MEMBERS
2	3.644579e-27	1.000000e+00	FLAG_MOBIL

FIG. 3. a) p-values table b) Corresponding Features selection out of the database.

B. Feature Engineer and Selection

After converting categorical variables to numerical ones and replacing categorical missing values with the most frequently appeared, we checked the distribution of each variable to decide whether or not it has skewness issues. We set the threshold to be 0.75. Some examples of skewed features are annual household income, number of children, family status, and housing conditions. We then take the log(feature +1) to make the distributions of the skewed features more normal. After the feature engineering step is done, we perform a univariate feature selection by computing chi-squared stats between each non-negative feature and class. We ranked the score of each value as below. We also calculate the p-value of each feature to check their statistical significance. As you can see from the table and the feature score visualization graph, the top five features which influence our potential target variable (loan status) the most are age, income type, education, gender, and whether or not they have a phone. The top five features are highly statistically significant even at 99% confidence intervals. We expect these five features influence our final results the most.

III. METHODOLOGY

We are interested in creating classes of applicants who are similar and want to check if these classes are particularly more likely to get approved or apply for loan for a particular type. As the data contains both categorical as well as numeric data we use a combination of 2 clustering algorithms k-means clustering for looking at numeric distance and k-mode clustering for looking at categorical data which uses hamming distance instead of euclidean distance. The combination of both these clustering methods is called k-prototype clustering which is being used for the current problem. To find the number of means in k means clustering we first choose the k which gives us the knee point in a cost vs number of clusters plot. Here the knee point is at k=5 and at k=2. But looking more closely at the clusters that got created for k=5 one can observe that the clusters are really close. Though the centroids are different but they are extremely close to one another. On looking at the centroids for 2 clusters one can see that the first

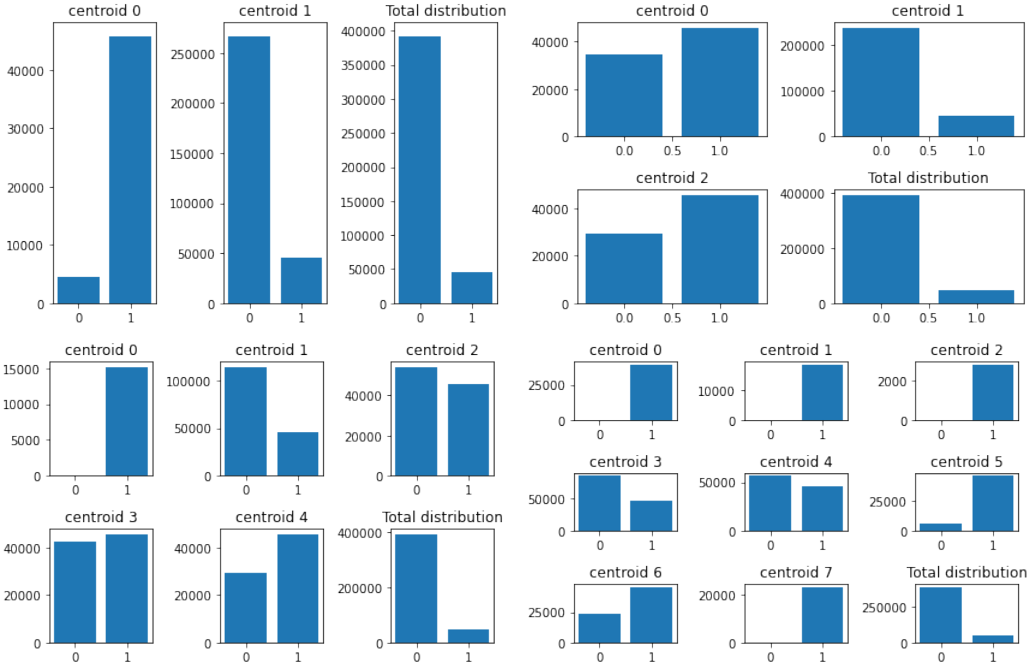


FIG. 4. Graphs of different Clusters.

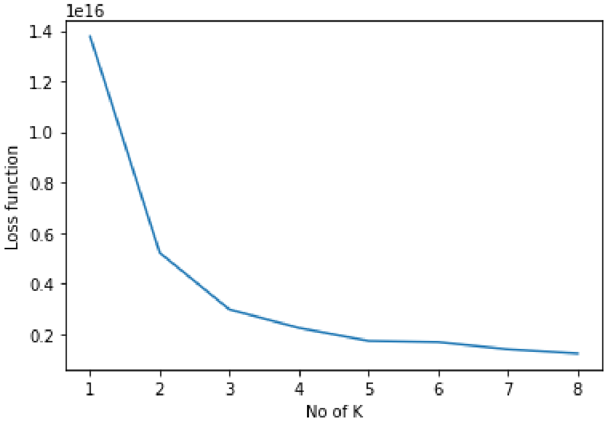


FIG. 5. Loss function w.r.t k

cluster divides the dataset into 2 distinct groups where one set will have high probability if getting approved while the second centroid nearly encompases all the points and is not really anything special. The centroid for points which are 1 is around. One of the strongest determining factors was found to be amount total income with income around 350000. That was something common among different clusters. Barring that all the clusters which were]approved had email accounts. When k gets increases the same trends in income and email still persist but one sees different clusters because of education level and type of employment. Ownership of cars, houses and phones seem to be having no effect on the clusters. While looking at demographics data the data could not be divided by clustering method. But behaviour data could divide the data into classes because it contains both the important axes. On weighting the parameters the clustering really did not change though the clustering now is more separated than in the earlier case.

IV. EVALUATION

CART and Logistic regression were the algorithms employed to select a subset of relevant features. We got the predictive accuracy of 0.42403 when we ran the Logistic regression on the original target Y variables of the STATUS

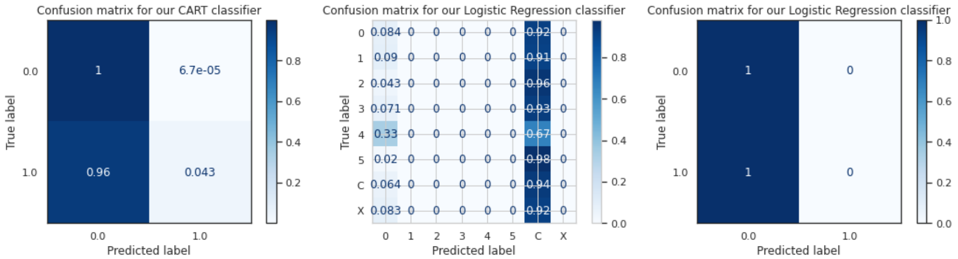


FIG. 6. Confusion matrices for different predictive models as labeled.

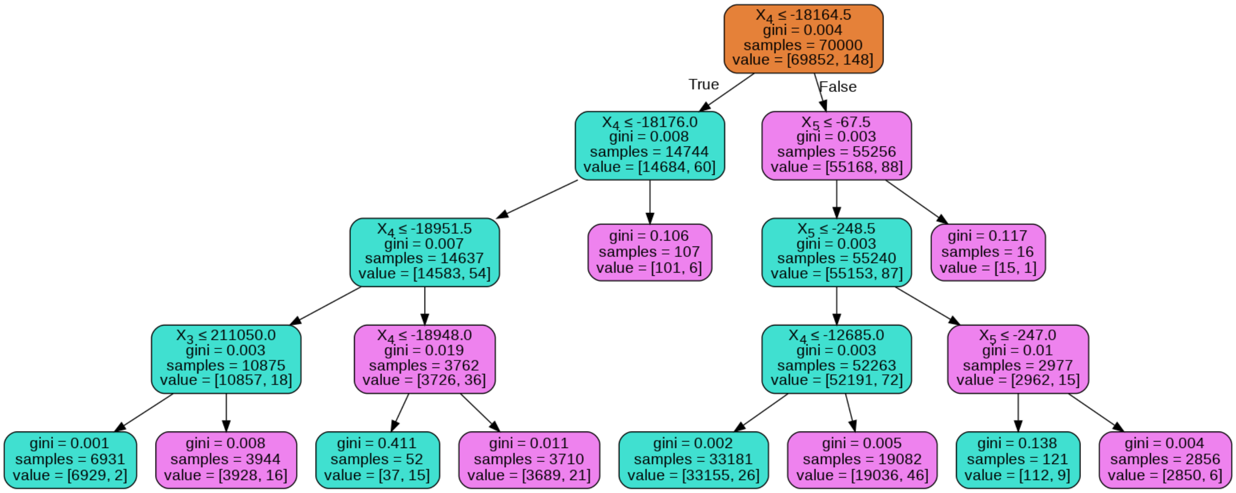


FIG. 7. Decision Tree classifier set to depth 4, with Regression implemented as in CART algorithm. The algorithm was run upto the depth 12 to find better accuracy but presented here for depth 4.

with 7 different categories, and then we divided the Target Y variables into 2 categories of likely or unlikely to get approved and our model's predictive accuracy got boosted over 99%. Similar was the case for CART, both these predictive models represented almost perfect classifiers over binary STATUS Target Y variables. The depth of the Decision tree classifier was set to 12, hence we got our supervised hierarchical segmentation model. In the figure we've shown till the depth 4 for the presentation purposes.

V. EXPERIMENT

Data Cleaning was implemented by replacing the NaN values of the occupation_staff feature with the mode value of the column. As the features dataset and target dataset files were of different dimensions, intersection of the two datasets was obtained and datasets were merged so that we could run our predictive model of Logistic Regression and Decision Tree to obtain the relevance of the features. Implementation of the correlation plot and pairwise plot of the 3 different features was done. It gave us insight into how correlation between certain features like total income earned, gender,ownership of the car was negative hence counter-intuitive. We observed no obvious pattern among these features. Hence it pushed us to develop the logistic regression predictive model using the elasnet algorithm which would enable us to recognize more relevant features.

Feature importance was calculated in the predictive models of CART and Logistic Regression. K-means and K-mode clustering was performed on our dataset as explained in Methodology.

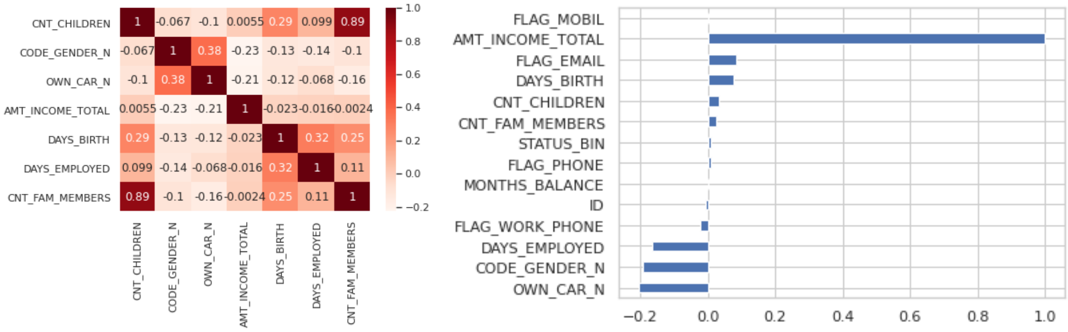


FIG. 8. Correlation graphs with respect to the feature *amount_income_total*

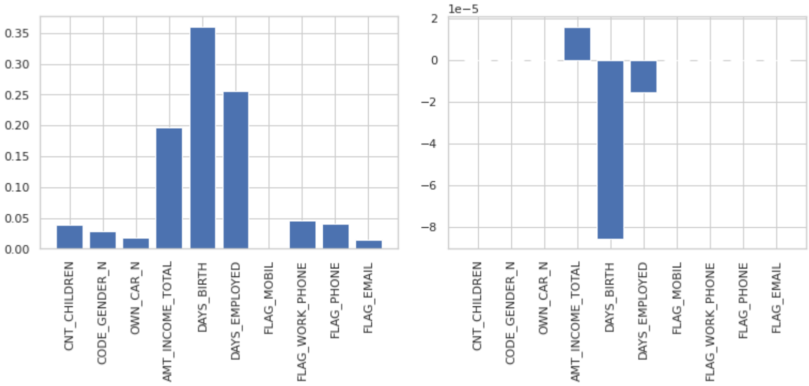


FIG. 9. a) CART feature relevance values b) Logistic Feature Relevance values 3 Features Total Income, Days of Birth and Days employed came out to be significant in both the models.Other features values were close to zero hence they contributed nothing to very insignificant amount.

VI. CONCLUSION

Some of the important features found from elastic net were found to be the features that get captured by k-prototype clustering. Income level was a really important feature that gets captured by k-prototype clustering. Number of days of work which did get captured in elastic net played no role in clustering. We have not been able to figure out which features are getting selected. On changing the scaling of different features, especially on the basis of feature importance it was found that, important features which are determining the clustering into classes 0 and 1 have not changed. Demographic data didn't allow us to cluster the data into relevant classes mainly because the relevant features were not present. Using behavioral or total data allowed us to cluster the data into relevant classes.