Business Problem

Lending club is an American peer to peer lending company. There has been a significant increase in the delinquency rate. The company would like to develop a business strategy in order to reduce the default rate. The dataset provided by the company consists of 300000 records and 151 features.

Credit Risk Modeling

A credit scoring model is used in evaluating a credit application. Credit lending firms can save millions of dollars by assessing an applicant's profile before approving a loan. The model estimates the probability of default using a machine learning algorithm. The model can assess the profiles of existing as well as the new clients. We will use logistic regression to predict the probability of default.

Logistic Regression is a probabilistic technique that uses a logit function for binary classification.

Logit = log(odds) = Bo+B1…Bn

We get, P = 1/1+exp(-y)

P = exp(Bo+B1…Bn)/1+exp(Bo+B1…Bn)

Where P: Probability of default

Bi: Regression coefficient of explanatory variables

Investigating Data Quality

Before going into predictive analytics, it is imperative to check the quality of the dataset. After reviewing the data quality, I found that the data quality is a bit concerning.

- **Proportion of missing values:**

There are incomplete values in the dataset. Nearly 90% of the features contain missing values and it is necessary to impute them. Records with <1% missing values can be dropped.

- **Proportion of outliers:**

"Annual_inc" consists of outliers that significantly affects its variability. Outliers can be detected by using a boxplot. These outliers could likely be a result of a numerical error.

- **Distribution of features:**

Some of the features are highly skewed and must be transformed before we proceed to advanced analytics. Skewness value of 0 implies that the feature is normally distributed. Generally, skewness value less than 1 is considered acceptable.

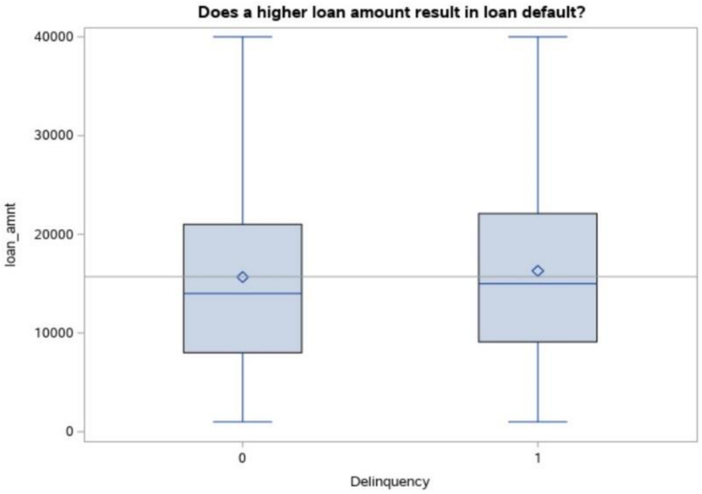
- **Multicollinearity between independent features:**

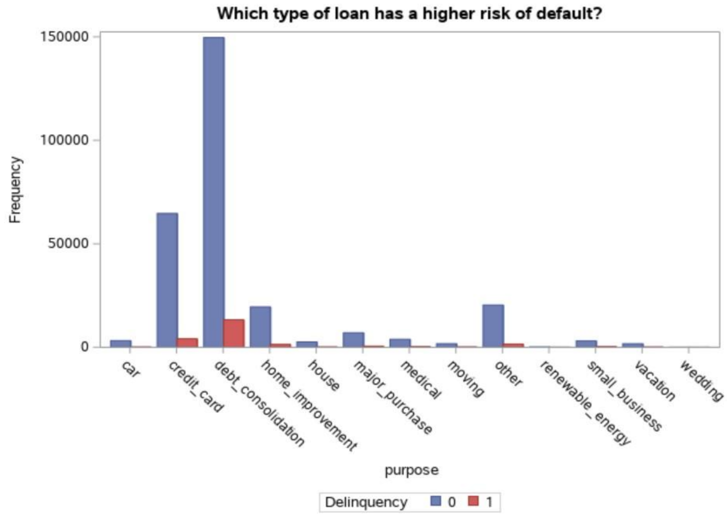
''Loan_amnt'' and "Installment" has a variance inflation factor of 21 and 10 respectively which implies both of them are correlated and one of them must be dropped from the final model. When there multicollinearity exists then we might get incorrect estimates.

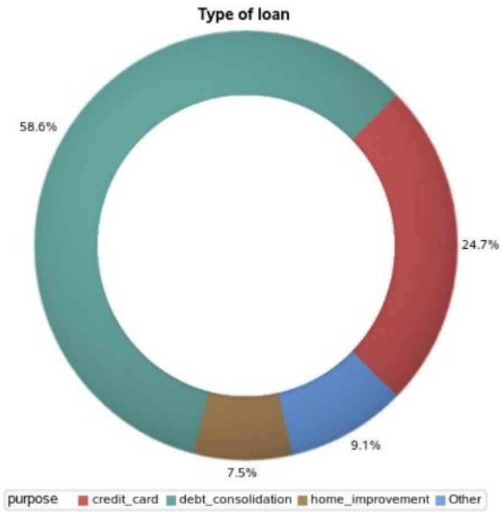
- **Class Imbalance:**

In our target variable, number of "No default" records are way more than the number of "default" records. Defaults are <10% of the target variable.

Exploratory Data Analysis

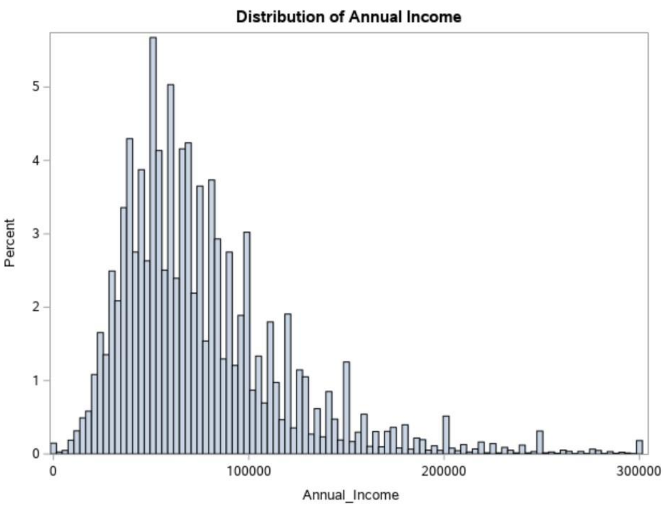


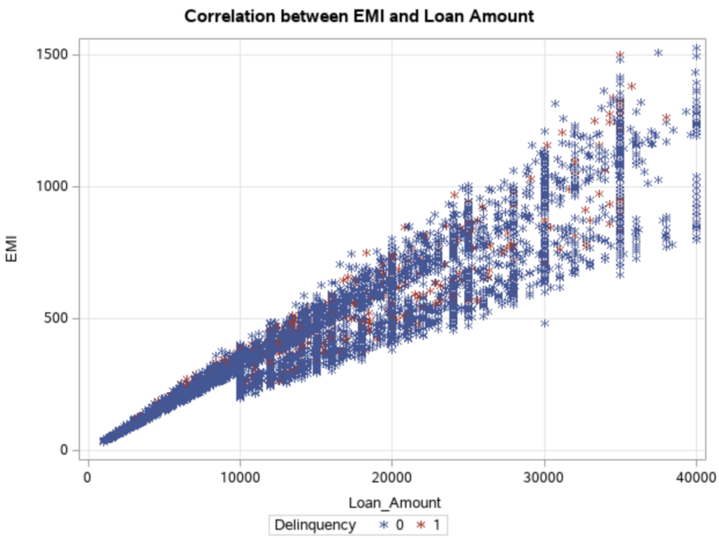




As it is evident from the bar graph, Debt consolidation and credit card have a higher risk compared to the other loan types. Interestingly, these loans are much more popular and a large share of company's revenue might depend on them. Debt consolidation loans and credit cards account for 83.3% of loans.

On the other hand, if we look at the first graph, average loan amount taken by clients who defaulted is almost similar to the clients who paid back on time. This shows that higher loan amount doesn't seem very helpful in identifying credit risk.





The histogram shows us the distribution of annual income and it seems that annual income is positively skewed. Most Customers get an annual income between $40,000 and $100,000.

Monthly Installment and loan amount are highly correlated. Since EMI is derived from loan amount, they have a high positive correlation between them. Also, Loan amount has a VIF of 20 which suggests high multicollinearity. This variable must be excluded from the final model.

Hypothesis Testing: ANOVA

Statistics is a field of study that deals with exploration and interpretation of numerical data. A hypothesis is an assumption about the population. We make an inference about the entire population using a sample of data. Analysis of variance or Anova is a statistical method to check if three or more groups are statistically different. For the study below, I will test the result at 5% level of significance.

Assumptions of Anova:
1, Normality
2. Homogeneity
3. Independent observations

Problem: To check if people with different annual income have equal fico score?

- Null hypothesis: There is no significant difference between the fico scores of people with different annual income.

  $\mu_{\text{Low income}} = \mu_{\text{Medium income}} = \mu_{\text{high income}} = \mu_{\text{very high income}}$

- Alternative hypothesis: There is a significant difference between the fico scores of people with different annual income.

  $\mu_{\text{Low income}} \neq \mu_{\text{Medium income}} \neq \mu_{\text{high income}} \neq \mu_{\text{very high income}}$

The GLM Procedure

Class Level Information

Class	Levels	Values
income_category	4	High Low Medium Very High

Number of Observations Read	300000
Number of Observations Used	299994

The GLM Procedure

Dependent Variable: fico_avg

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1068993.0	356331.0	290.61	<.0001
Error	299990	367830519.6	1226.1		
Corrected Total	299993	368899512.6			

R-Square	Coeff Var	Root MSE	fico_avg Mean
0.002898	4.973039	35.01632	704.1232

Source	DF	Type I SS	Mean Square	F Value	Pr > F
income_category	3	1068992.997	356330.999	290.61	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
income_category	3	1068992.997	356330.999	290.61	<.0001

For our Anova test, we got a p-value<0.05. Null hypothesis is rejected as we have insufficient evidence to support our claim. Thus, Fico scores are significantly different for customers belonging to different income category.

Possible reason: Customers who make low or medium level income might have faced financial problems due to which they weren't able to pay back the loan before due date. This resulted in a reduction of their fico score. Also, high credit utilization could have impacted their fico score.

Model Development

For the process of building a Binary classifier, logistic regression has been used. Logistic regression predicts the probability of default where (p=0) would be taken as “no default” and (p=1) would be taken as “default”. Cutoff value of the probability would depend on company’s risk preference.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.0623	2.9937	0.1259	0.7227
term	36 months	1	-2.4415	0.0402	3686.2890	<.0001
int_rate		1	-0.6782	0.0398	290.3893	<.0001
installment		1	3.7613	0.0439	7356.0584	<.0001
grade	A	1	-2.9988	0.2142	195.9851	<.0001
grade	B	1	-2.3450	0.1899	152.5207	<.0001
grade	C	1	-1.8289	0.1684	117.9038	<.0001
grade	D	1	-1.2744	0.1451	77.1527	<.0001
grade	E	1	-0.7213	0.1301	30.7374	<.0001
grade	F	1	-0.1550	0.1312	1.3971	0.2372
annual_inc		1	-0.0705	0.0233	9.1877	0.0024
verification_status	Not Verified	1	-0.1801	0.0286	39.5839	<.0001
verification_status	Source Verified	1	-0.1203	0.0261	21.2427	<.0001
purpose	car	1	0.4987	2.9895	0.0278	0.8675
purpose	credit_card	1	0.8065	2.9876	0.0729	0.7872
purpose	debt_consolidation	1	0.8705	2.9875	0.0849	0.7708
purpose	home_improvement	1	0.7160	2.9877	0.0574	0.8106
purpose	house	1	0.6265	2.9895	0.0439	0.8340
purpose	major_purchase	1	0.7878	2.9881	0.0695	0.7921
purpose	medical	1	1.0362	2.9885	0.1202	0.7288
purpose	moving	1	0.8612	2.9894	0.0830	0.7733
purpose	other	1	0.7603	2.9877	0.0648	0.7991
purpose	renewable_energy	1	1.5397	3.0042	0.2627	0.6083
purpose	small_business	1	0.8349	2.9887	0.0780	0.7800
purpose	vacation	1	0.8881	2.9898	0.0882	0.7664
application_type	Individual	1	-0.3481	0.0365	90.9601	<.0001
delinq_2yrs		1	0.0240	0.00892	7.2259	0.0072
out_prncp		1	-3.3066	0.0420	6196.5252	<.0001
total_pymnt		1	-2.6223	0.0347	5703.9228	<.0001
last_pymnt_amnt		1	-2.9909	0.1176	646.2911	<.0001
income_category	High	1	-0.0229	0.0458	0.2497	0.6173
income_category	Low	1	-0.0834	0.0309	7.2720	0.0070
fico_avg		1	0.4204	0.0155	731.7093	<.0001
last_fico_avg		1	-1.8509	0.0141	17160.3877	<.0001

Relevant features are selected using backward feature selection method. This selection method creates the best possible model with significant features. The above table shows all those relevant features that can impact the target variable. In logistic regression, Wald’s chi-square test is used because the dependent variable is categorical. All levels of purpose, grade_F and income_category_High are not statistically significant as their p-values are more than 0.05.

Model Performance: ROC Curve

ROC stands for “Receiver Operating Characteristic”. The performance of all possible cut-off values is included in an ROC curve. An ROC curve graphically summarizes the tradeoff between true positives and true negatives. True positive rate (sensitivity) lies on the y-axis and False negative rate (1-specificity) lies on the x-axis. The threshold value or cut-off point can be determined using J-statistic. For our final model, we got an auroc score of 0.94.

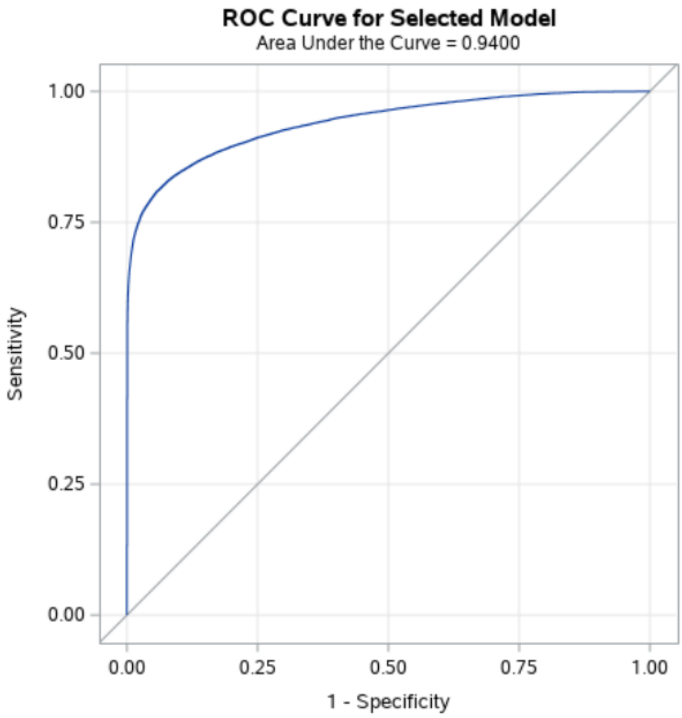
The LOGISTIC Procedure

Model Information	
Data Set	WORK.LOAN_DATA
Response Variable	Delinquency
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	296996
Number of Observations Used	296996

Response Profile		
Ordered Value	Delinquency	Total Frequency
1	0	272102
2	1	24894

Probability modeled is Delinquency=1.



Insights for Effective Business Strategy

- Clients with annual income of more than $95000 have a higher fico score and are less likely to default.
- More than 83% Clients have applied either for a credit card or a debt consolidation loan.
- Risk of default doesn’t depend on the amount of loan. A person with a small amount of loan might find it difficult to settle his debt.
- Clients with lower interest rate generally pay their installments on time.
- A high monthly installment can impact the probability of default. In credit industry, debt to income ratio of 25% is considered acceptable.
- Loan term and delinquency are negatively related. Clients who have applied for loans with a tenure of 60 months or higher are less likely to default.
- Delinquencies are mostly linked to individual accounts. A Joint loan application is much more reliable as a co-applicant brings additional source of income and other assets.

| Bona-fide work of Aarush Gandhi |

EBEON1120326850