Introduction:

Loans Default, predicting whether the loans that may be defaulted or not. As I am a part of industry that provides loans, it is very important to atleast have 98% of recovery rate. Then only it would consider as a profit which further improves/expands the business. It can be predicted through data sciences because, we have all the information on a customer available and we can build predictive models based on that.

Summarize the problem statement you addressed.

The problem statement you addressed:

We are predicting the default rates of a loan as it is a significant part of lending business because lenders must predict whether giving out a loan will result in profit or loss. Normally, loans are profitable because of interest, but sometimes a borrower will default, which will be a loss for business. Thus, it is important that the lender is able to gauge the likelihood of a borrower defaulting before making a loan to him/her.

Summarize how you addressed this problem statement (the data used and the methodology employed, including a recommendation for a model that could be implemented).

How you addressed this problem statement:

For this project I have gathered three different datasets,

1.) Customer Loan Details over years from a bank
2.) Unemployment rate over the years in United States of America
3.) Natural Disasters Data across United States of America

All the three data sets are combined into a single data set that has all the available information for a model to predict the outcome.

Used logistic regression model to identify the relation between variable loanDefault based on the variables interest rate, Fico Score, Annual Income, Unemployment rate during the loan defaulted year, Any disaster during that year.

loanDefaultModel <- glm(loanDefault ~ interest rate + Fico Score + Annual Income + Unemployment rate + disaster , data = lendingData_df, family = binomial())

Summarize the interesting insights that your analysis provided.

Analysis:

When summary applied on the model it implied that Unemployment rate, interest rate and Annual income had the highest impact on the loan default.

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.655e+01	2.400e+03	-0.007	0.99450

INTEREST RATE	1.368e+00	4.868e-01	2.811	0.00494 **
FICO SCORE	1.418E+01	2.400E+03	0.006	0.99528
ANNUAL INCOME	1.653e+00	6.094e-01	2.713	0.00668 **
UNEMPLOYMENT RATE	1.084e+00	4.990e-01	2.172	0.02984 *
DISASTER	4.089E-01	2.673E+03	0.000	0.99988

Fico Score and Disaster during that year did not have much impact on the loan defaulting when compared with other variables. When did further analysis of why Fico score didn't impact the defaulters, because Fico score itself is calculated based on different parameters.

Summarize the implications to the consumer (target audience) of your analysis.

Implications:

Interest Rate and Annual Income are playing a major role on the loans that are getting charged off as per the datasets that we used for analysis. Higher Interest Rate results in high monthly payments which are tough to repay if their Annual income is less or if they have any other financial commitments. So, when we provide a loan, we have to factor the annual income of a customer and some other financial details related to customer based on which we can approve the loan amount (either decrease or approve the same amount applied for).

Discuss the limitations of your analysis and how you, or someone else, could improve or build on it.

Limitations:

One observation I was expecting based on the datasets or my assumption was Disaster also would play an important role because; I am working in the industry which deals with loans, COVID19 pandemic incurred losses to so many sectors due to which so many lost their jobs and were unable to repay the monthly installments. To help customers there were so many disaster programs were offered on the loans like zero interest for 3-6 months, clearing the past due amount till they got relief because of pandemic.

To improve the model, I should have also considered the disaster programs that provided by governments or financial institutions which helped the customers during bad phase during all the years to repay their loans.

Concluding Remarks:

Current model predicts that the interest rate and annual income are significant variables that decides the outcome of the loan whether to become a charged off or not. Also the accuracy of the model was 74.45% which implies not a bad model with the data that we have at hand. There are variables like Fico score and Unemployment rate which did not have a significant impact on the model. To conclude Higher the interest and Lower Annual Income would increase the risk of loan getting charged off that could impact the business of a financial institution.