finalProject2_KanaparthiVenkata

Venkata Kanaparthi

5/23/2021

How to import and clean my data

Data Import

We can use read_csv() to import the data into a data frame Ex: lendClubLoans_df <- read_csv(“data/finalProject/Final_Pro
stringsAsFactors = FALSE)

Data Cleaning

1.) Removed Duplicates 2.) Check for a person whether he has multiple loans are not. But did not find.

Fields that are retained from the original data set are which I feel can be utilized to predict the loan status

id
member_id
loan_amnt
funded_amnt funded_amnt_inv term
int_rate
installment home_ownership
annual_inc
verification_status issue_d loan_status addr_state
fico_range_low
fico_range_high

What does the final data set look like?

```
## # A tibble: 6 x 16
##        id member_id loan_amnt funded_amnt funded_amnt_inv term      int_rate
##     <dbl>     <dbl>     <dbl>       <dbl>           <dbl> <chr>     <chr>
## 1 1077501   1296599      5000        5000            4975 36 months 10.65%
## 2 1077430   1314167      2500        2500            2500 60 months 15.27%
## 3 1077175   1313524      2400        2400            2400 36 months 15.96%
## 4 1076863   1277178     10000       10000           10000 36 months 13.49%
## 5 1075358   1311748      3000        3000            3000 60 months 12.69%
## 6 1075269   1311441      5000        5000            5000 36 months 7.90%
## # ... with 9 more variables: installment <dbl>, home_ownership <chr>,
## #   annual_inc <dbl>, verification_status <chr>, issue_d <chr>,
## #   loan_status <chr>, addr_state <chr>, fico_range_low <dbl>,
## #   fico_range_high <dbl>
```

What information is not self-evident?

Data looks good. No Issue

What are different ways you could look at this data?

1.) Interest Rate Vs Loan Status (Charged off or not) 2.) Annual Income Vs Loan Status (Charged off or not)

How do you plan to slice and dice the data?

#Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain. We have multiple status of loans in the loan status field, would like to take only customers those were charged off into one variable and others into another variable of data frame type.

How could you summarize your data to answer key questions?

We use summary() function on the model that is used which helps us in providing the R and Rsquared if Linear model and AIC, Deviations if it is a Logistic regression. It will provide us the information on how predictors influence the outcome

What types of plots and tables will help you to illustrate the findings to your questions?

Scatter plots, histograms, line graphs

Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

Would like to use logistic Regression on the data, As Charge off Yes or No(Categorical Variable) is the outcome based on the predictors (Continuous or Categorical variables).