Inferential Statistics

The section on inferential statistics is looking at the statistical significance of observations made and thoughts had during the EDA. This is an essential step to understanding whether or not the differences between the customers who subscribed to term deposit and the customers who did not, are factual or just happened by chance.

The focus for us lies in three categories:

- Hypothesis testing for continuous features.
- Analysis of discrete features.
- Collinearity and Heatmap between features.

Hypothesis testing

The hypothesis testing is done for continuous variables to check if the differences to the target variable ‘y’ caused by the independent features are statistically significant or not. A T-test is used for hypothesis testing since the data does not represent the whole population.

Three variables that are analyzed are:

- Age - age of the customer
- Balance - balance in the customer’s account
- Duration - call duration of the latest contact.

A hypothesis to determine if the differences within the independent features that lead to customers subscribing to a term deposit or not are statistically significant. The p-value for all three features is less than 10^-5 which allows us to reject the null hypothesis.

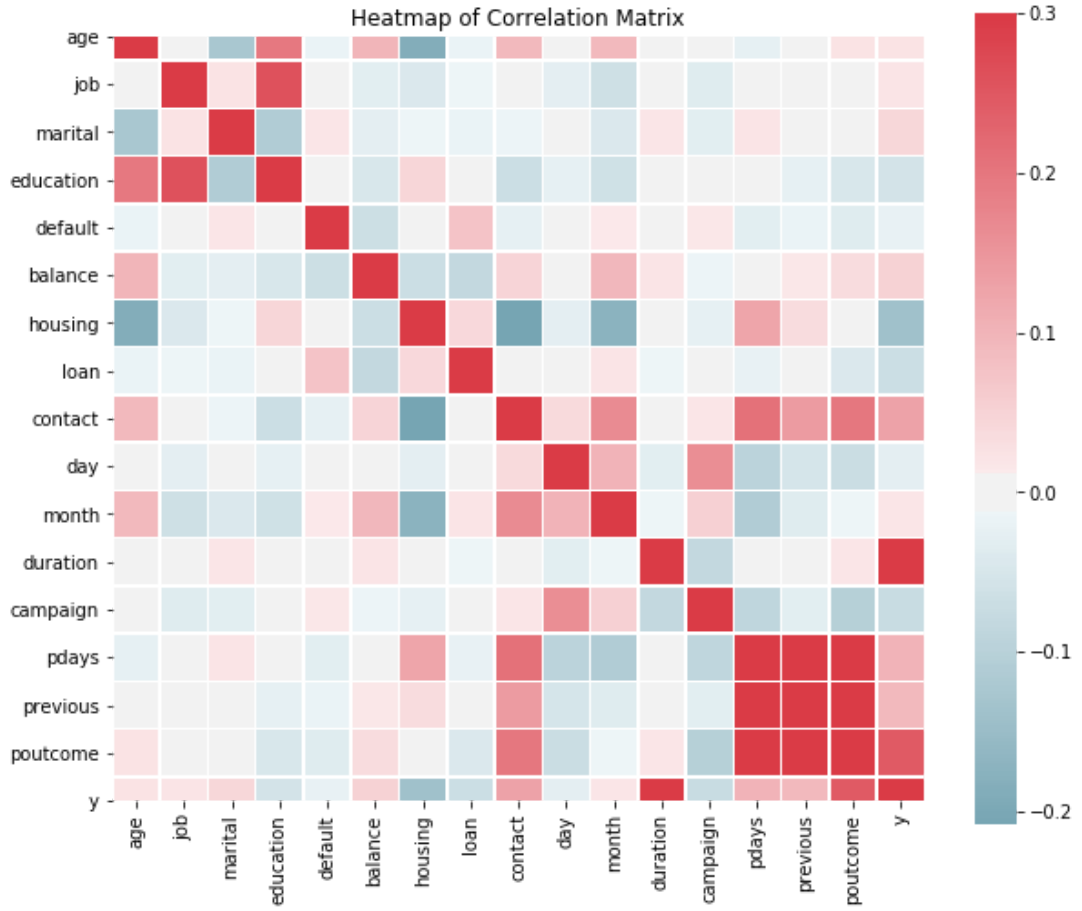
There was one more hypothesis test done on the balance in a customer’s account. The Null hypothesis was to determine whether the customers with zero or negative balance who subscribed to term deposits are similar to customers with a positive balance. The Null hypothesis was rejected with a p-value of 10^-300.

Having a positive balance definitely leads to higher customers subscribing to a term deposit.

Analysis of discrete features

A normalized stacked bar plot is plotted for various discrete variables to check each independent feature's effect on the target variable.

Collinearity and Heatmap between features



The target feature 'y' has a positive correlation with the continuous variables age, balance, and duration(duration is strongly correlated).

The values of discrete variables are converted manually from string to numerical quantities.

The strong correlation can also be seen with the 'poutcome' feature as well. But the data has a very high number of unknown values.

Conclusion

The data definitely proves that the dependent variables that cause the differences in the target variables are not by chance.