ML Capstone Project1

# Credit Card Approval

1.Why is your proposal important in today’s world? How predicting a good client is worthy for a bank?

There are so many factors have impact on credit card approval.

In this credit card data set we have Ind id, GENDER, Car Owner, Property Owner, CHILDREN, Annual income, Type Income, EDUCATION, Marital status, Housing type, Birthday count, Employed days, Mobile phone, Work Phone, Phone, EMAIL\_ID, Type\_Occupation, Family members, label.

Here target variable is label.it has 1 & 0’s.it means 0 for the accept the credit card approval, 1 for the reject credit card approval.so it is classification model and it has binary classification (1&0).so I select the logistic, random forest, decision tree, KNN models. these are the best models for classification.

We have so many columns. we need to clean data.it means checking null values, fill null values, checking for the outliers and caping them, check the data distribution, convert categorical columns to numerical columns, checking correlation, split data into dependent variable and independent variable, check dependent variable data distribution, transform the data if needed. all columns should not have high impact on target variable. we can select only some of the features which are having high impact on target variable. It is classification logistic regression method so in this we can use logit method (p>0.05) to select best features which are having high impact on target variable.

Now need to split the data into train and test and perform standardization if needed(scaling).

Create the model and pass the train data into model to understand the data pattern. then pass the test data to predict the values for passed test data. Compare the test data and predicted data how correctly the model predicted the values.by using confusion matrix.

2.How is it going to impact the banking sector?

My proposal is very much needful in banking sector. because who are giving approval for a credit card need to know about all issues. ex. age, salary, kids, income so many factors. If you sit and discuss about all things it can take lots of time. But if we can use these methods we can easily got to know that which person is the right person to get credit card or not. No need to concentrate all features, only some features are needed. All techniques are in this model to got to know which person get credit card or not very quickly. based on this models easily can take decision about credit card approval.in future if use these models less time taken, all computer work, no formalities to filling forms like that.

3. If any, what is the gap in the knowledge or how your proposed method can be helpful if required in future for any bank in India.

Yes,I felt difficult in some stages in doing this model.

1.to understand some features like birthday count, working days.here need to convert data in to years. After taken so much time I have used divide with 365.its worked. Also had negative values. but finally resolved this issue.

2.another difficulty was tried to transform dependent variable. Because its completely biased about 0.data is not distributed normally. i tried to do transform.

I used all methods in transform which are I known like log, sqrt, pow transformations.log transformation not suitable this one. the situation got worst. sqrt transformation is nothing changed. in pow transformation also tried boxcox it is not applicable because data has 0 values. its worked on positive values only. I faced so much difficult to use yeo jhonson transform.so I just leave it again tried. My model got good accuracy. So I left this problem.

In this credit card data set we have both categorical and numerical columns. and also target column is in another data set. based on common column I merged these two datasets. there is null values in some columns. Type of occupation had highest null values. i filled it is unknown occupation constant word. i filled these null values with that column median if it is numerical column. Birthday count, employed years these are in days. i changed these into years by divide with365.

Annual income, employed years, children, family members have outliers.

By using IQR method I caped those outliers.

Converted categorical data into numerical data by label encoding.

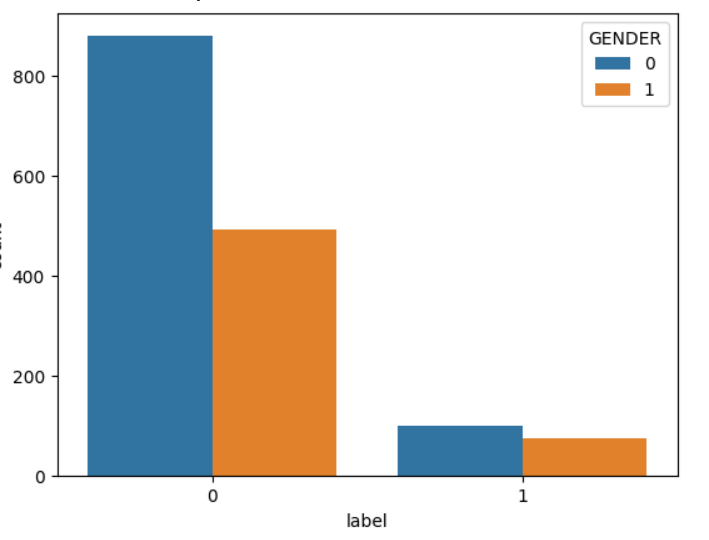
I checked all columns still any columns have outliers like that. no.

I checked the correlation . I did data analysis.

## My observation on data:

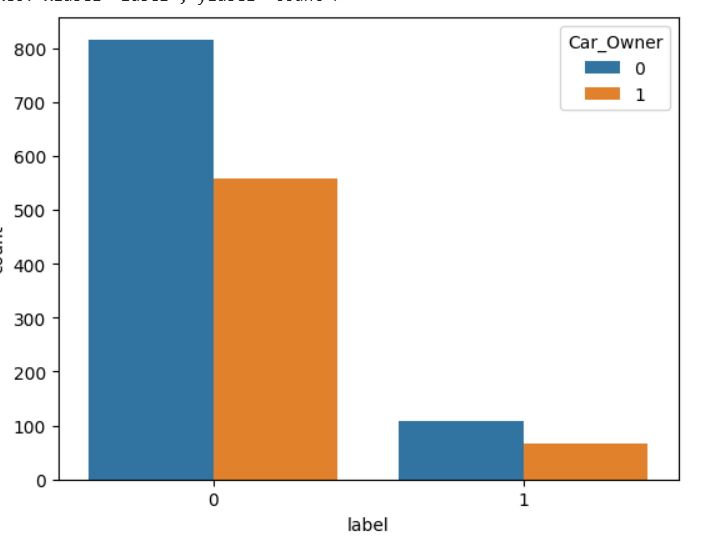
What I observed on this credit card data is who got high approvals on credit card Who got less approvals, what type of profession people got high approvals and less approvals all analysis is

In gender who got high approvals, less approvals

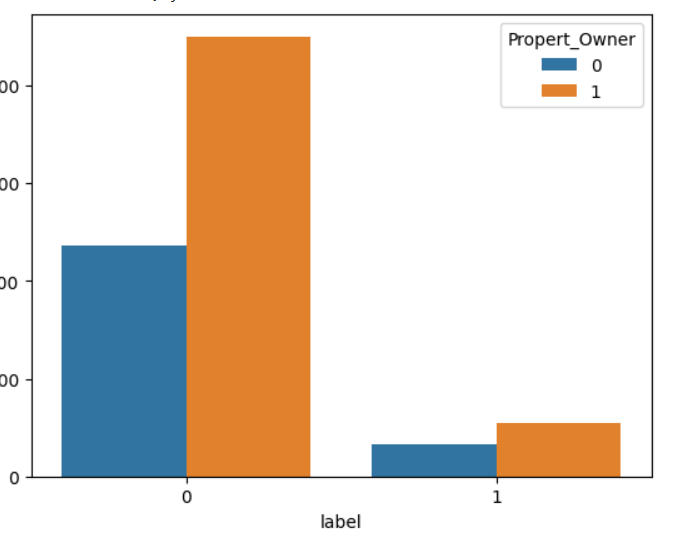


here females got highest approvals for creditcards.0 for female,1 for males

who got high approvals, having own car people or not having own car people



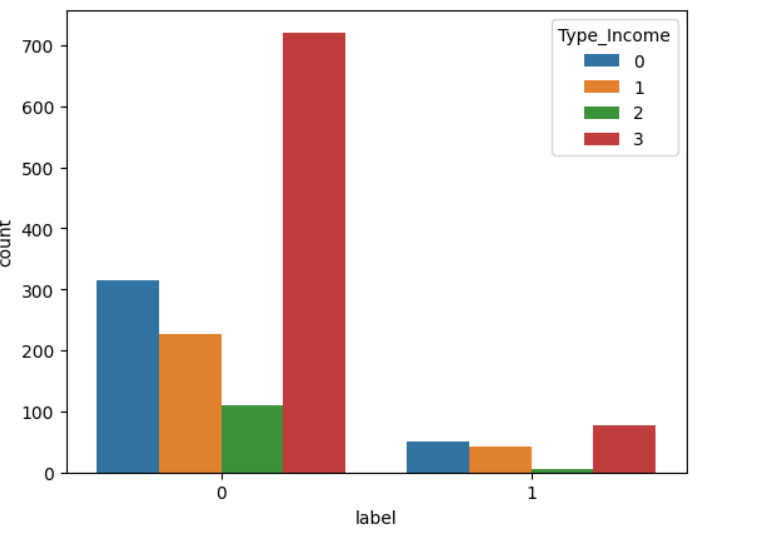
who are not having cars they got high credit cards. who having cars they got low approvals for creditcards.0 for not having cars,1 for having cars.



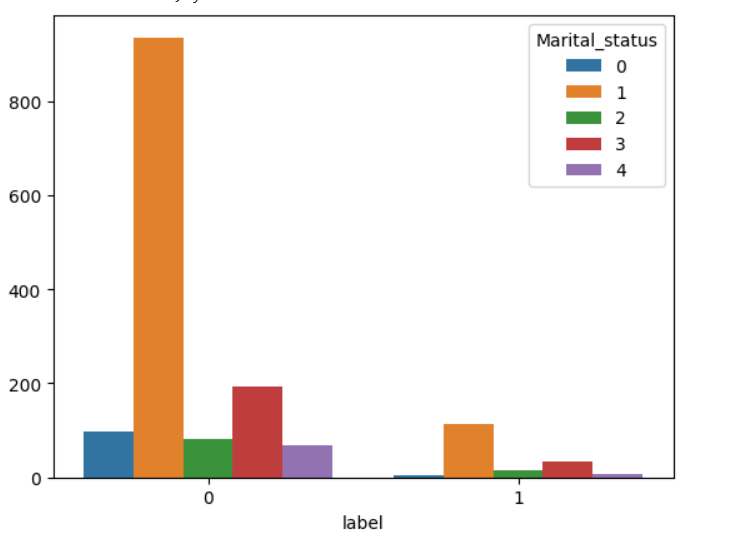
who have own property they are highly preferred to get credit card approval.1 for having own peoperty,0 for not having property

who does not have own property they are less preferred to get credit card approval

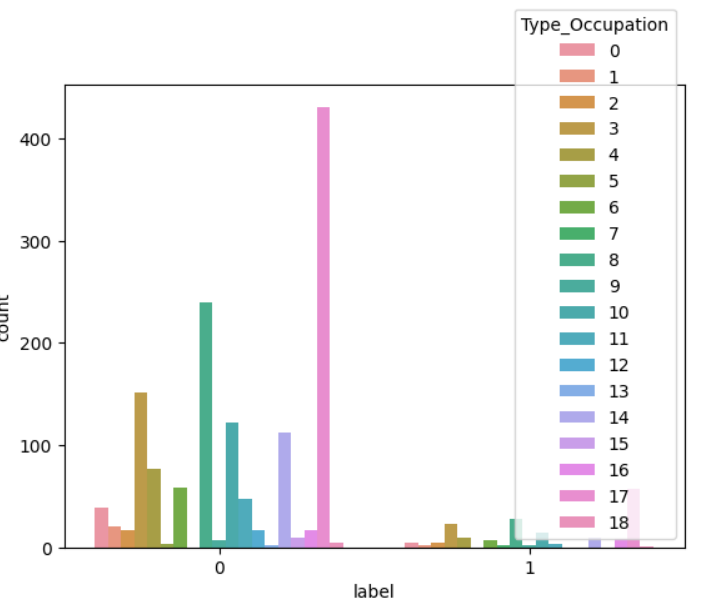
Which type of income people have got highest approvals of credit cards



who are working they got high approvals for credit card, next commercial associate got high credit card approvals, next pensioner also got some approvals for credit cards, but who are state servent they got low approvals low rejections of credit cards



Here who got married they got high approvals of credit cards. after this who not get married or singles are got high approvals. remaining all got very low approvals for credit cards. widow got very less approvals for credit cards.



Unknown occupation people got highest credit card approvals, next

Laborers got second highest approvals for credit cards. next

Core staff reality agents, HR staffs got very less approvals of credit cards

after analysis I did split the data as x and y.

here I used logit method to feature selection. based on p>z values I removed columns which are have >0.05 value.

At last I selected index id, GENDER, Type Income, Marital status, Type Occupation Employed years features.

Again I split the data as x train, y train, x test, y test.

Scaling independent data by using standard scalar method.

I created a model logistic regression because it is a binary classification target column. passed the data x train, y train through this model for learning data pattern. and passed the x test data to predict the values.

Import the confusion matrix to test how your model working. i used it and I got 89% of model accuracy. To improve this I used another models also.

I checked random forest model also. with this model I got 90% of model accuracy I got.

I checked decision tree model also. with this I got 86% of model accuracy I got.

I checked KNN model. with this model I got 89% of model accuracy.

In these four models random forest model got highest model accuracy which is 90%.