SOCIAL NETWORKING ADS

ABSTRACT

The Dataset contains information about users on a Social Networking site and using that information as features for our ML model, the model predicts whether a particular user after clicking on an ad on the Social networking site goes on to buy a particular product or not. It is a CLASSIFICATION PROBLEM as the output says whether the user buys the product or not, so it's either a yes or a no. Well this particular Social Network has a Business client which lets assume is a car company which advertises itself by putting adds on the social networking site. Now the work of the social network here is to gather information as to whether the user bought the product or not.

PROPOSED METHOD

The dataset used is the social_networking_ads imported from :
https://www.kaggle.com/rakeshrau/social-network-ads .
The data set has information about the User ID, gender, age ,estimated salary of many such users and whether they purchased the product afer clicking on the ad or not.
The dependent variable in this case is Purchased which is 1 if user purchases the car and 0 otherwise. So to predict whether the user bought or not the goal is to create a classifier which would put each user into the correct category .

The following features will be considered as the independent variables :
1)Age
2)Estimated Salary
The dataset also contains 3 more columns and we are leaving them because:
1) The UserId has no effect on whether the user would purchase the Car or not
2)Some might say that Gender would play a role but that is really subjective to discuss.

I have used feature scaling on the data used as features for prediction to achieve better accuracy and have used 3 classification algorithms to see which gives the best accuracy :

1) Logistic Regression
2) K-NN classification
3) RandomForest classification

RESULTS

I found K-NN model gives the best accuracy for this data set with :
- Train accuracy: 90.67 %
- Test accuracy: 93.0 %

The Logistic regression model gave train accuracy and test accuracy of 82.33% and 88.0% respectively while the random forest classification model gave train a ccuracy and test accuracy of 98.0% and 90.0% respectively. Though the random forest gave better accuracy but we also see overfitting.

CONCLUSION

The social networking ads dataset is a classification problem and I achieved maximum accuracy using KNN-classification algorithm of train accuracy and test accuracy as 90.67 % and 93.0 % respectively.