SECTION 1

We used a dataset which contained 20000 rows and 27 columns where the rows represent the data about 20000 different people of India and the columns represent the data about their spendings and earning and 25 more parameters overall it’s a very bid data set …….

SECTION 2

The approaches done by us is we tried using 4 different models and then choose one to go on with first we worked on the entire raw dataset without cleaning or filtering the data the four models which we made are Linear Regression, Decision Tree Regressor, Random Forest Regressor, Support Vector Regressor (SVR), Gradient Boosting Regressor (GBR) and XGBoost Regressor. Above 6 models were tried by us we found out the performance by calculating the MSE and R square of all the models

SECTION 3

We compared all these model and we choose the Linear Regression for our further analysis as our data contains a lot of attributes out of which some are highly co-related it was difficult to choose the predictor variable and In the beginning we even got a perfect model that is R square value equal to 1 in linear regression because of overfitting.Also for the remaining models some of them gave us a very good R square score between 85-99% but it is more or less due to the high co-relation in our dataset because of which we are getting such a good R square value. So first we had to deal with this problem we have used two different approaches to solve this problem First approach was to get rid of the attributes using domain specific knowledge of our dataset and second approach is finding the co-relation between the predictor variables with the predicted variable and getting rid of all the attributes which have low co-relation .The main problem lies in cleaning our dataset as along with 20000 different rows we have 27 attributes out of which 25 are integers and most of them have very high co-relation with each other that is why we even got a perfect fit model due to this high co-relation there was overfitting in our model and we ran into a problem of multi-collinearity . Even PCA can be used to get rid of extra attributes and only choose the attributes which are important as PCA is a dimension reduction technique one more algorithm which is used to reduce the nuber of dimensions is LASSO.So after comparing our two approaches reducing dimension by domain specific knowledge and reducing dimension by finding the co-relation we found that ……… approach is better because ……

SECTION 4

The results of our selected approach can be seen below we have calculated the MSE and R square of our model we have even made a graph to calculate the expected v actual prediction and we can even see the residual plots of our approach .Now in this approach first of all we have reduced the number of attributes in our dataset this was very important because of high co-relation in almost all of our columns in the data frame our predictions very going very wrong even though the value of R square and MSE was coming good for our model but the reason for that to come good was actually the high co-relation in our dataset and not the performance of our model

SECTION 5

By seeing the output of our selected approach we can conclude by saying that our model can give accurate predictions …% of times. This specific approach is good because of the kind of dataset we were dealing with ….

SECTION 6