Final Report

***Abstract*—*I have applied 5 Machine Learning models on three datasets related to different domains. The first dataset is about Credit card fraud detection, in which principal component analyses has already been done because its confidential data related to financial accounts, as this is a classification problem I have applied SVM with various kernels and Naïve Bayes machine learning algorithm. Dataset has about 300 thousand records with imbalanced classification so used various sampling techniques. Some SVM kernel performed better than others. The second dataset is about crop production where I was predicting the production based on Area for different districts using linear regression and regression tree ML methods, this dataset has a lot of categorical variables therefore dummy coding has been used. Regression model was found more satisfactory as compare to linear one. The third dataset is about online store/shop transactions of certain period which I have analyzed and applied clustering technique (unsupervised learning) for predicting consumer behavior based on together purchased items.*

I. Introduction

The objectives of this poject is to apply various machine learning models on three large datasets. The reason behind choosing datasets are different for each dataset.

- Crop Production: Food is one of the fundamental neccesity for everyone on this planet.With the increasing number of population the demand for more food is also increasing therefore agriculture plays an important role. It would be beneficial for farmers and help them in taking future decisions.I have analysed previous years data crop agriculture dataset of INDIA , and trying to predict production.The dataset for this analyses is available at

  https://data.gov.in/resources/district-wise-season-wise-crop-production-statistics-1997

- Credit Card Transactions: People do online transactions from one euro to lakhs of euros every seconds. Sometimes the payment system become vulnerable and can be misused by someone. This fraud transactions costs a lot to individual, to company and as well as to government not just in terms of money but also in terms mental peace.The best way to deal with it is to extract out those transactions somehow by building some system which can help in reducing these kind of fraud in a cost effective way which does not involve any legal consequences. Here I have analysed the data and build a model to predict the fraud transactions among all the transactions. I was evaluated the accuracy of this model using confusion matrices. The dataset for this analyses is available at

  https://data.world/vlad/credit-card-fraud-detection

- Online Retail Dataset : Market trends has always changes from time to time. It is important to know what customer wants and how easily a store can provide it (be it a physical store or an online store).The price , availablity and how the items are buying by the customer tells about their behaviuor and shop owner can use it to increase their sales and can mange the store in a better way.Similary I have taken the data of an UK online store and by using clustering techniques I have tried to predict consumer behavior for example prediction of sales and segmentation of items. The dataset for this analyses is available at

  https://archive.ics.uci.edu/ml/datasets/Online%20Retail

II. Related Work

A. Crop Production Dataset

As Liakos et al.[1] explains how machine learning can be used in agriculture segment which is refered as digital agriculture. Data is very important and useful in building such models which can bring revolutionary changes in the field. In this research the main focus was on crop management , water management , soil management and livestock production. It explained that how machine learning works for supervised and unsupervised leaning. Yield prediciton has been achieved by ML methods, the aim is to apply cost effective and low price model for predicting coffee fruits on a branch which helps coffee growers in their economic state which is similar to my work as I am trying to predict yield for vaious crops on the basis of area.They have also used regression tree model.

[2] V.shah and P.shah also used Multiple Linear regression model for predicting ground nut for few states in their research. The good thing is that they have considered other environmental factors such as rain , biotic factors like ph value, and Area in the model building ,unlike the model I have used where I would be majorly depend on the Area factor and that is the limitation of dataset I have used.

S. Mishra et al.[3] says in the similar research model build on the prediction of “corn” crop using ANN (decision tree) and muliple linear regression model. They have compared both the model in process. Regression analyses done on the dataset and showed that corn production is not dependent on evironemnt factor instead more on a planting practices , which is very useful information.Correlation of dependent and independent variables has been also established using statistical analyses.Time series analyses has also be done to compare the previous year data and train the model accordingly.

B. Sitienei et al.[4], also studied on Tea crop production in the kenya region, as tea is produced in around 58 countries also for maize. Studies has done for creating multiple regression model for wheat yield in the past, similar approcah has been taken here for Tea.Climate variables considered here and contigency table used for verification of the model.Several statistical model has been used for showing trend analysis like t-test , correlation analyses and multiple linear regression analyses. Results has been shown seasons wise (rainfally) which can be very useful for taking further decisions related

to tea production.[5] K. Olson and G. Olson has observed about creating a model using soil properties and climate factors. Here also multiple linear regression has been used to generate and analyze coefficients , also time series data has been taken of corn crop of last 43 years to build the linear equation.

[6] Ramakrishna has done similar work in their research by applying multiple linear regression and regression trees.They have build a model on randomly selected data.Evaluating the model by calucluating MPE values , where linear regression underperform with high biasness.The next model used was regression tree which performed well.

[7] In this research crop production model build for corn, which is one of the valuable crop.The data collected is of US time series data with remote sensing data.Some specific US states has been target for this.The goal was to produced contry level prediciton , so they performed aggreation and build yield model on that.Resuls showed the forecasting season wise.

B. Credit Card fraud Detection Dataset

[8] In this research thay have been try to resolve challenges related to fraud detection methods and second about new methods for intrusion detection , money laundering , spam detection etc. E-commerce made this complusary to look after credit card transactions to find fraud detection by finding trends and suspicious transactions. SVM ML method is used as similar to my model, for this supervised learning. Neural network has also used. Multiple supervised learning models has used to build hybrid fraud detection system.Positively we can say that hybrid system are much efficient as multiple techniques involved , but on the other side complication has also increased to make such models on similar datasets.

[9] W. Lim et.al. Explains how to build find fraud detection system by using Transaction aggregation method ,which calculates sum of transaction amount respective to each feature in last few days. Every transaction is labeled as “fraudulent” and “legitimate” ,which is the dependent variable.Weighting system has been used in this. The positve aspects is the accuracy of the mathematical model they form but the downside is the complication of weighted column which is not easy to interpret and execute.

A. Thennakoon et.al[10] has mentioned in the research how to chose a suitable algorithm for fraud patterns and also compare various machine learning algorithm for the same.In the dataset used in this reasearch, credit card no is being hashed due to security reasons.Also the dataset is highly imbalanced like the one I am using in my research.Data cleaning , transformation, normalisation and reduction has been done prior to getting a final dataset. For handling the imbalanced classification SMOTE and CNN sampling techniques used and also 10-fold cross validation already applied before resampling. They have used SVM,Naïve bayes ,KNN and logistic classificaiton model,and then they made confusion matrices for the performance of the best model. We can observe the Positive aspects of this reasearch because of its process in creating models for fraud detection and the accuracy they have achieved showed in the paper.

F. Carcillo et al.[11] suggested in their research paper that instead of using supervised learning methods which is widely use for fraud detection , they have used hybrid of supervised and unsupervised learning model , which can handle the anomalies and performed better by using outliers detection at different levels.It says that the combinations was really efficient but there was a lack in the accuracy of the prediction.However in the research.

N. Carneiro et.al[12] suggested that focus of the research should be on more practical scenarios from the perspective of e-commerce merchant.They have also explore automative and manual classier also the process of selecting and labelling of data.They have not just do statistical analysis but also conducted intervies with fraud analyst to understand the fundamanetals and to identify a pattern of fraud transactions.They transformned and trained data on support vector machine, logistic regression and random forests ml methods.They have taken the key indicators of a transaction like automation level , chargeback , rate of payments refused and processing speed.They have calculated the suspician score and accepting/rejecting a transaction on the basis of a threshold decided. All the three applied supervised learnings gives good result. So far this is best work I have find which takes care of e-merchant also.

C. Online Store Dataset

R. Gupta and C. Pathak[13] explains in this research paper , they have been trying to predict sales based on dynamic pricing.They have done the EDA and do the customer grouping by using k means clustering approach. They have followed widely accepted RFM technique and applied logistic regression method to find out which items are more likelable to buy by customer.The positive aspect is that they build the model by keeping customer and organization both in their mind.

[14] F. Weber and R. Schütte have explain all the challenges faced in Retailing and machine learning initiatives which can considered and categorized.It has been observed that mostly retailers wants to predict their sales.They have show all the ML techniques which can be used as consolidated summary with the increased complexity by adding more products in the cateogory , competetion and store location.This article is good to understand overall concept about online store and ML methods used to predicting sales and customer behaviour.

[15] In this research , used supervised learning to predict customer behaviour and identificaiton of the items bought frequently.This comes under the Market basket analyses and they have used Apriori and Prim algorithm. It is important to know this behaviour so that shop owner can change their strategies accordingly. It is very easy to do online camplaigns and getting feedback which can also help in building the model.They have done deep comparison between both apriori and prim and found that both are suitable depending on the situation.One big aspects is Apriori is good for big datasets and prim for continuos data. They have applied association rules on the subset of the data , not on the complete dataset.In general this process is much efficient , they have both the approached which can handle multiple situation.

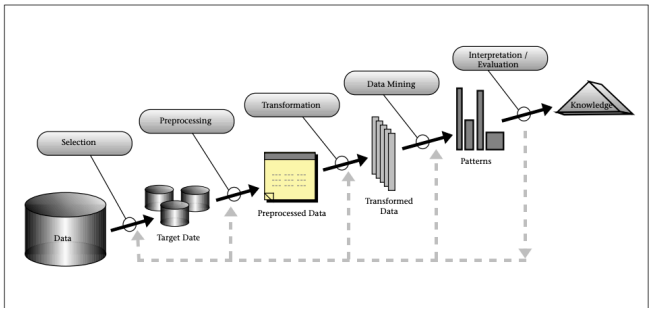
As R. Liu et.al[16] explains in this work, they have segmented two classes VIP and no VIP thorugh RFM technique (I am also using RFM in my model), basically to keep the relationship with the customer , it is required to understand their needs. For that they analysed their transaction data , they

used k-means clustering and association rules to assign an item into a particular cluster.After they done Market basket analyses to understand customer behaviour and created a record. The positive aspect is that they also suggested how items should be placed like men wear should be one click away from women wear. One drawback can be noticed is the history information of the organisation is very shot to produce meaningful information.

III. DATA MINING AND METHODOLOGY

Fayyad et al.[20] Knowledge discovery in databases(KDD) is one of the emerging field and most demanding methodogy in current time. For making any system efficient we need meaningful data , as data can we available in any form with lots of noise and biasness therefore we need to manually analyse it and pre precessed it for further analyses .KDD is one such approach which can be use here. Some sectors has even require it more like healthcare management , defence management in which accuracy plays a very big role. More the data is noise free it is easy to use it.

Process of KDD



For all the dataset I have used the KDD methodology. Data selection, preprecossing then transformation ,using Data mining and evaluation different models and interepreted the results.

A. Credit Card fraud Detection Dataset

For credit card dataset , PCA has been done already because users credit card information and other user details are confidential.As this is the classification problem of supervised learning.
Dependent Variable : Class which has two values False and True.

Data Preprocessing :

- Import the data avaiable in csv into R studio.Load the neccesary libraries like dplyr for data manipulation, caret and catools for data splitting.
- Check the head and structure of the datasets which says there are total 32 variables and mostly all of them are numeric.As PCA has done already , there is no need to perform normalisation.
- There was no Missing values in the dataset.
- Naïve Bayes Method : It is a machine learning classified which used maximum posterior probablity rule for classification. It works on the past events data to predict the future possiblities based on the bayesian method.
- Support Vector machine : Is a classified is non-probablistic classifier used in classification as well as regression.It used kernel for internal mapping which mapped high dimentional fetures with its input.I have used vanilladot and rbfdot kernels in this research.
- Dataset has contained 284807 , records which is very huge. Before applying ml methods, I have binning my dataframe, changed all numerical variable into cateogorical bins of same length. All the colums are factors with level 5 now , and it is easy to run the model. OneR package has been used which gives the Bin method.
- As this is a classification problem I have checked for any imbalanced classification in my dependent variable and found that out of 284807 , approx 99.9% data falls into False(non-fraudulent) and only remaining on the fraud data. I have split my data in to training and test data using caTools library split function as 80:20 ratio,and run my model on this. After that I have predict my train model against complete test data which gives 99.9% accuracy (TP+TN / total number of caes) means 56961 cases has been correctly predicited which shows the overfitting of the model.
- To tackle imbalanced classification I have used two techniques: SMOTE and downsample
- SMOTE-sampling is used for handling unbalanced classifcication.Two parameters perc.over and perc.under used to control the over sampling and under samplaning of majority and minority classes. This is the midway.N. Chawla et al.[17] explains that it is the combination of up sampling and downsampling which gives the hybrid sampled data and handled unbalanced.Smote sampling is provide by DMwR package in R.
- C. Tantithamthavorn et al.[18] explains Downsample is about make the majority class frequency same as minority class.It basically downsampled the majority class frequency. It is an R function downSample provided by caret package.I can not use up-sampling as it will add frequency in the minority class same as majority class resulting in about 4 hundred thousands records which might be very difficult to run.
- I have run SVM first on SMOTE sampled data using kernel “vanilladot” which is linear kerner and majorly used for sparse data , but here I am using it for comparing it efficiency with “rbfdot” which is more generic kernel based on Gaussian radial basis. Then I run SVM using downsampled training data using vanilla dot.
- After that ,I have train the model using naïve bayes on downsampled data , as smote data is still too large for naïve bayes.

- I have predict the above model against the original splitted Test data.
- We compare the evaluaton in evaluation method.

Data Modelling :

I have chose Naïve bayes and support vector machine algortihms as they are widely used for problems like these.

B. Crop Production

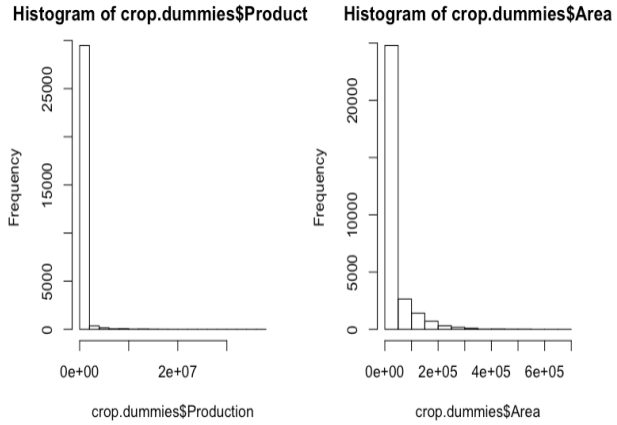
I have majorly taken 4 crops Rice,Banana ,Turmeric and Sugarcane.

Dependent Variable : Production (continious data).

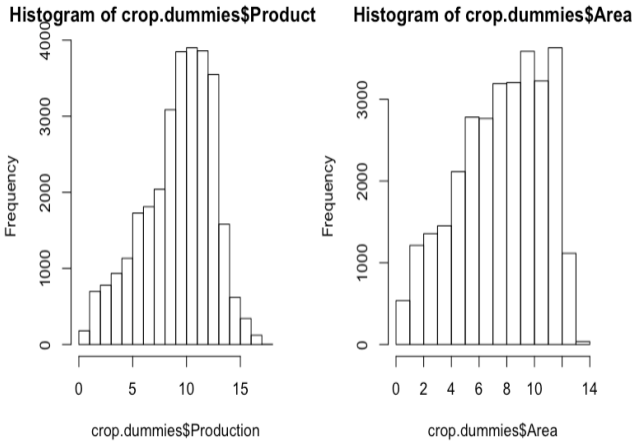
Data Preprocessing :

- Import csv file into R. It contains 7 variables and 246091 records.
- Check the Na values and omit them.
- I have checked the coorelation using Ggally and coorplot which are R functions. I have Area field which is highly correlated with my dependent variable.
- Structure of the dataset shows that there are lot of categorical variables for District name and Crop which si creating the problem while prediction.To deal with problem I have created dummy values using dummyVar R function which have do internel encoding of categorical variables into 0 and 1.
- After that I have checked the skewness of Area and Production continious variable which are very high (11 , 3.33) and making the model inefficient , I have applied loglp on both variable and skewness reduced to (-0.5 ,-0.3).

Before : Skewness



After Skewness



- After that I have split the training and test data into 70:30 ratio and train simple regression model and made predictions. Draw various plots of residuals which we disucss in the evaluation part.
- Then I run regression tree model and plot the model fitness using rpart.

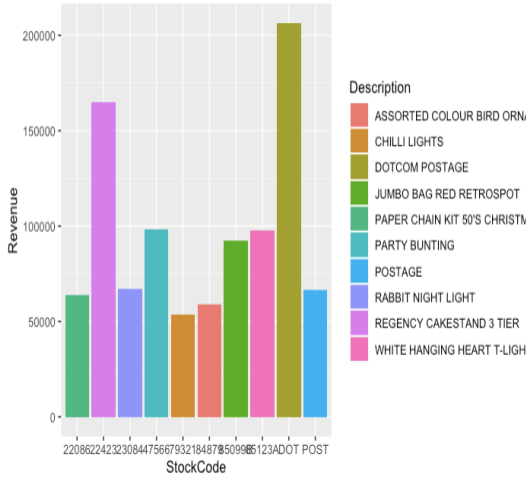
Data Modelling:

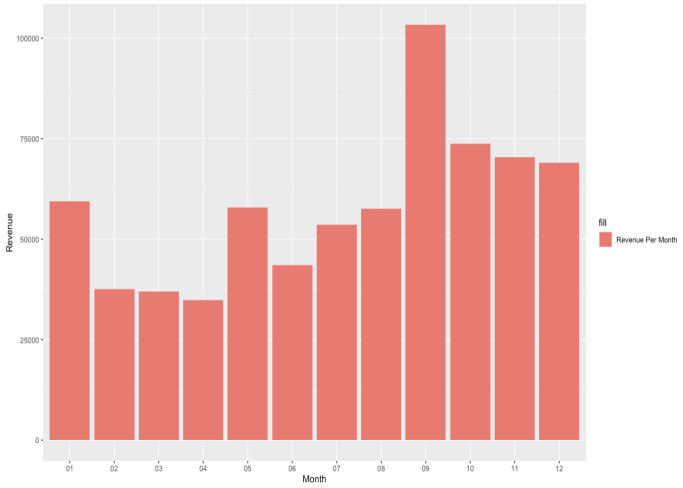
I am buidling a model for crop production using muliple linear regression and regression trees.

C. Online Store

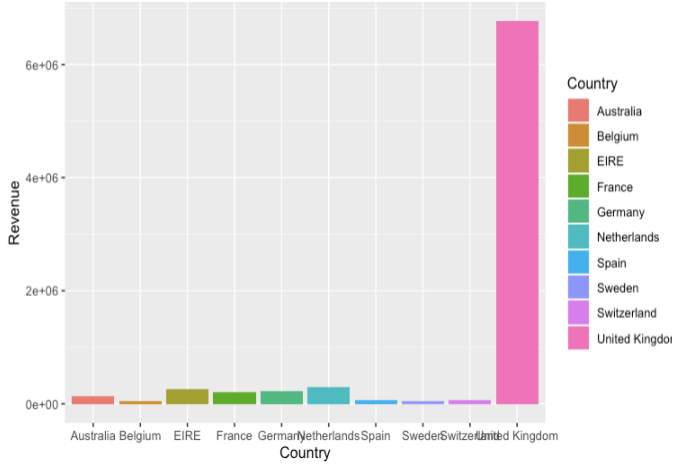
Data Pre-processing:

- Import xlsx file into R.
- Convert the data type of invoice date into a date format using anytime R function.
- Converted invoiceNo, StockCode, Description, and Country variables to factor for making it categorical and avoid any other values. Simultaneously converted Quantity and Customer id into integer from numeric to avoid float and double values.
- Now colSums gives the NA values column wise, 9288 values in customer id are null means invoice has not been generated for them. We cannot remove them if we want to find top ten selling items by the store. So added a new column revenue (quantity*unit price), for finding top ten items. I have showed the top ten items and see **Dotcom Postage** is the most selling item of the store.

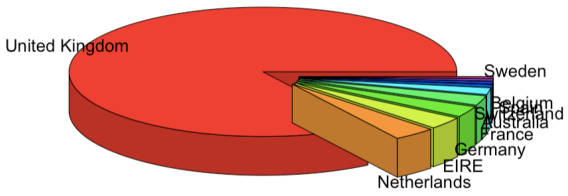




- Now we can remove NA values from customer ID, and filter it records for above the above top 10 selling items.
- I have added month for visualising the transactions per month.
- I have showing the top 10 countries which produce high revenue. We found highest revenue producer are the people from UK who bought items.



Pie Chart of Top 5 Countries



- I have also showed the revenue per month and see maximum sales occurred in September month.

Data preparation for clustering:
I have created clusters by grouping customer id and country columns. Calculated customer regularity by checking the transaction for unique year and months for every customer. As data was for 13 months so divide it by 13 which give me the regularity rate. Then created new items for every stock in the new data set for clustering. Then I have performed the PCA as I have very wide dataset.S. Mishra et al[19] PCA is method to analyze a data of inter correlated dependent variables.Internally it worked on Eigen values and eigen vectors.

Data Modelling :

I have created clusters and apply k-means and also one regression model just to see the outcomes. I discussed its limitaiton and effectiveness in the evaluation part.

IV. EVALUATION

After succefully run all the models for three datasets. Let's explain model evaluation and parameters.

A. Credit Card fraud Detection

For evaluation I have used confusion matrices and calculating the accuracy of the model. Confusion matrices is good measure for classification models.

Confusion Matrix for Model using Vanilladot svm kernel

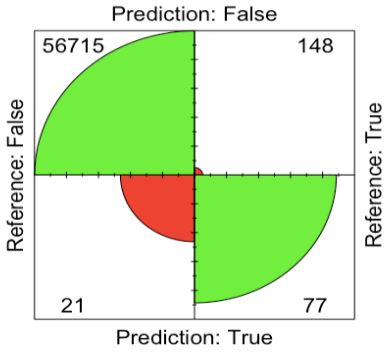


TABLE I. CONFUSION MATRIX FOR VANILLADOT SVM KERNEL

Confusion Matrix	*Positive*	*Negative*
Positive	56715 (TP)	148 (FN)
Negative	21(FP)	77(TN)

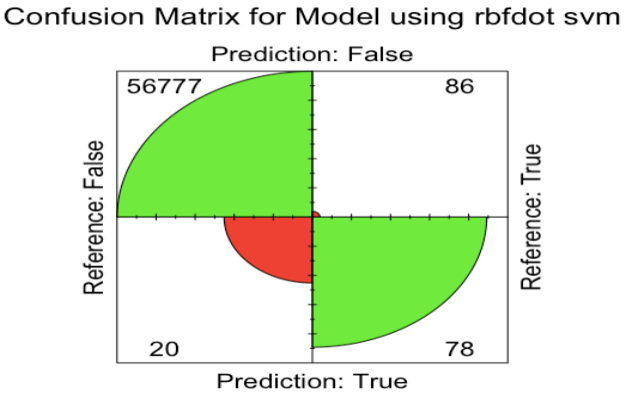


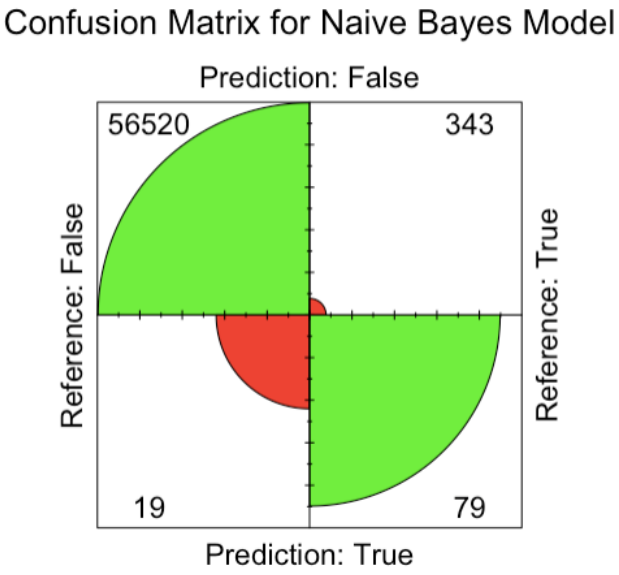
TABLE II. CONFUSION MATRIX FOR RBFDOT SVM KERNEL

Confusion Matrix	*Positive*	*Negative*
Positive	56777 (TP)	86 (FN)
Negative	20(FP)	78(TN)

By looking at above matrix we can say vanilladot svm has correctly predict non fraudulent transction for 56715 cases, and incorrectly do false prediction for 148 into fraud transaction which suppose to be in positive category. Similarly it has incorrectly do false prediciton of 21 cases and correctly for negative 77 cases.

If we talk about accuracy (TP+TN / total number of cases) which is 99.7% , specificity is (TN/FP+TN) 0.78 (0.0 is worse and 1 is perfect) and for rdfdot svm accuracy is 99.82% and specifcity is 0.78 almost similar. .We can say that svm vanilladot performed slightly better than rbfdot kernel.

Now evaluate accuracy for naïve bayes model. We can see in the below accuracy is 99.3. which is not bad against the whole test data. Also I have observed that naïve bayes took only few seconds to run smote sampled data which have total one hundred thousand rows.



Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision
0.9996639	0.1872038	0.9939680	0.8061224	0.9939680
Recall	F1	Prevalence	Detection Rate	Detection Prevalence
0.9996639	0.9968078	0.9925914	0.9922579	0.9982795
Balanced Accuracy				
0.5934339				

B. Crop Producion

For multiple linear regression , the summary is

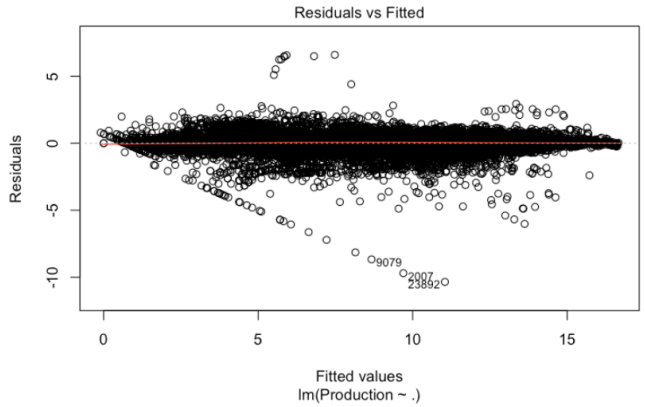
```
Residual standard error: 0.6307 on 20508 degrees of freedom
Multiple R-squared:  0.9635,    Adjusted R-squared:  0.9623
F-statistic: 845.3 on 640 and 20508 DF,  p-value: < 2.2e-16
```

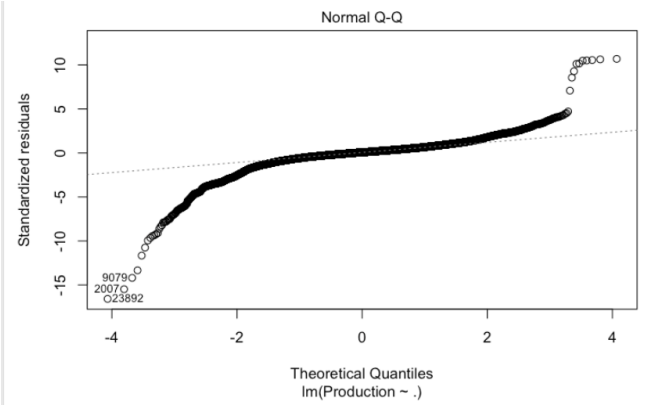
we can see after reducing the skewness the R square is become 96 % , it means the variance in Production can be predicted from independent variables by 96%. F value is very high (845) means that all my independent variables are not statistically significant.

	residuals_LM <dbl>
1	1.5310981
2	-0.8509352
4	1.6726017
5	-2.8416988
6	1.3660737
7	-0.3562010

6 rows

Residuals basically tells about the difference between the observed valus and fitted values means how much variance can not be explained by our model.



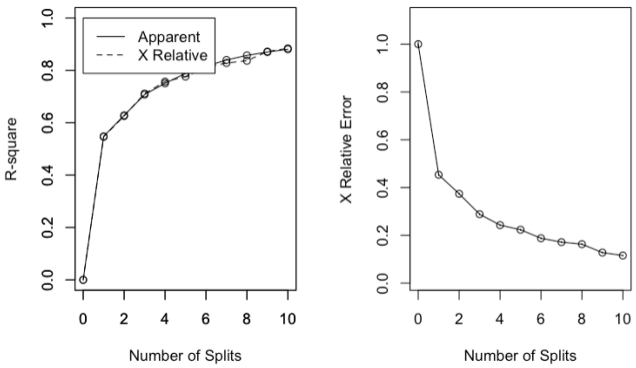


Normal QQ plot is saying that the two quartiles are coming from normal distribution if it is making almost a straight regression line else not. We can see that this model is underperform and not best model for this data. Also the value of MAPE (mean absolute percentage error) is "inf" means it in neagative.

Lets see regression tree plot

```
Regression tree:
rpart(formula = Production ~ ., data = train.LMData)

Variables actually used in tree construction:
[1] Area           Crop.Banana    Crop.Sugarcane Crop.Turmeric

Root node error: 223389/21149 = 10.563

n= 21149

        CP nsplit rel error  xerror      xstd
1  0.548127      0   1.00000 1.00003 0.0091700
2  0.080310      1   0.45187 0.45351 0.0039900
3  0.079289      2   0.37156 0.37406 0.0034760
4  0.043050      3   0.29227 0.28849 0.0027881
5  0.037498      4   0.24922 0.24293 0.0025069
6  0.027810      5   0.21173 0.22352 0.0024208
7  0.023758      6   0.18392 0.18792 0.0022871
8  0.017486      7   0.16016 0.17168 0.0022735
9  0.012949      8   0.14267 0.16305 0.0022236
10 0.010606      9   0.12972 0.12788 0.0019547
```

Root node error is 10 which is not good but also not that much bad.



R-square is almot 0.9 (90%) the variance in Production can be predicted from independent variables by 90%.

C. Online Production

I have calculated the pca percetage for each dataset , PC for first point is 15% and so on.

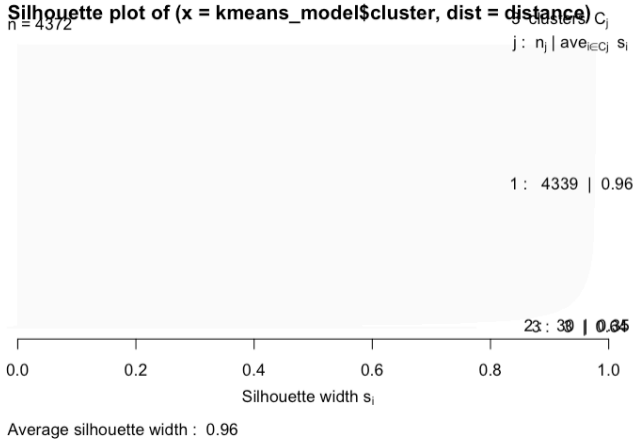
```
# PCA using prcomp():
pcaOutput = prcomp(retail.data2, scale = TRUE)
pcaVar = pcaOutput$sdev^2
PcaPercentage = pcaVar / sum(pcaVar) * 100 # in percentage
PcaPercentage
```

```
 [1] 15.531455 10.847035 10.698673  8.859074  8.393440  7.628211  7.425306  7.103443  6.502776
[10]  5.630631  4.733086  4.547457  2.099413
```

I have considered regularity , cutomer id , net revenue for top selling stock items , and chose centers 3(k value) as I have total 13 columns.

```
  1    2    3
 30 4339    3
```

Most of the data points clustered around cluster 2.Then checked the silhouette plot which telles how much the similarity of cluster points in their own cluster as compare to other clusters.



its value is 0.96 (should be in the range of -1 ot +1) , says that object value is matched mostly with its own cluster.

V. CONCLUSIONS AND FUTURE WORK

A. Credit card Fraud Detection

Credit card fraud detection is very popular studies for researcher because its scope is very large and the pettern of frauds keep changing over the period of time. In my research both the supervised learning works very well. There was not much difference in the accuracy while running svm with different kernels but overall “vanilladot” kernel works slighty faster than generic “rbfdot”. But naïve bayes model runs very quickly as compare to svm and almost same accuracy has found in the evaluation against the complete test data as svm methods.If we talk about the limitations we could say as this was not the real dataset , its already pca data , there is lack in predicting real time fraud for financial transactions. Also as I have not the expert , I believe in future if we can work with domain analyst and take real time feedback , we can make the system much better.Also if we can include more predictors like geo location and device information then a very efficient model can be created.

B. Crop Production

As crop is one the basic neccesacity of a human being, scientist around the world studies this topic regularyly.I have run multiple linear regression which underperformed with high biasness and one of the reason is involvement of high number of categorical variables with multiple levels. I Regression tree model run faster as compare to linear. There are not many predictor variables other than area is one of the limitation of this work. For the future we can add additional features in the dataset and predict the crop produciton with taking geo satellite variables and environment variables.

C. Online Production

There is a lot of scope in this dataset , we can either only visualise customer behaviour using EDA process and then we can also predict the sales using regression and clustering algorithms. There are some limitations inmy proposed work like I have not analyze the clusters thorugh visualisations. Clustering algorithm can be implemented in more mature way by anaysed it more thoroughly. With the limited knowledge of clustering algorithms I have not achieved much in prediction and training the model but I thorouhly done the overall analyse of the customer behavior through various visualisations. I will be looking forward to extend this work for more understanding of clustering algorithms.

References

[1] K. Liakos, P. Busato, D. Moshou, S. Pearson and D. Bochtis, "Machine Learning in Agriculture: A Review", *Sensors*, vol. 18, no. 8, 2018. Available: https://www.mdpi.com/1424-8220/18/8/2674/htm. [Accessed 13 December 2019].

[2] V. Shah and P. Shah, "Groundnut Crop Yield Prediction Using Machine Learning Techniques", vol. 3, no. 5, 2018. Available: https://www.researchgate.net/publication/326112319_Groundnut_Crop_Yield_Prediction_Using_Machine_Learning_Techniques. [Accessed 13 December 2019].

[3] S. Mishra, D. Mishra and G. Santra, "Applications of Machine Learning Techniques in Agricultural Crop Production: A Review Paper", *Indian Journal of Science and Technology*, vol. 9, no. 38, 2016. Available: http://www.indjst.org/index.php/indjst/article/viewFile/95032/74105. [Accessed 13 December 2019].

[4] B. Sitienei, S. Juma and E. Opere, "On the Use of Regression Models to Predict Tea Crop Yield Responses to Climate Change: A Case of Nandi East, Sub-County of Nandi County, Kenya", *Climate*, vol. 5, no. 3, 2017. Available: https://www.researchgate.net/publication/318496719_On_the_Use_of_Regression_Models_to_Predict_Tea_Crop_Yield_Responses_to_Climate_Change_A_Case_of_Nandi_East_Sub-County_of_Nandi_County_Kenya. [Accessed 13 December 2019].

[5] K. Olson and G. Olson, "Use of multiple regression analysis to estimate average corn yields using selected soils and climatic data", *Agricultural Systems*, vol. 20, no. 2, pp. 105-120, 1986. Available: https://www.sciencedirect.com/science/article/pii/0308521X86900624?via%3Dihub. [Accessed 14 December 2019].

[6] Ramakrishna Anil, A statistical approach to estimate seasonal crop production in India, Department of Computer Science, University of Southern California, Los Angeles,CA.[online].Available: https://pdfs.semanticscholar.org/5c74/91174c1d8b2029a8b273c6e04a95aa77cea4.pdf [Accessed on: Nov 2,2019]

[7] Y. Cai et al., "Crop yield predictions - high resolution statistical model for intra-season forecasts applied to corn in the US", *Gro Intelligence*, 2019. Available: https://gro-intelligence.com/yield-model-pdf/us-corn. [Accessed 14 December 2019].

[8] C. PHUA, V. LEE, K. SMITH and R. GAYLER, "Comprehensive Survey of Data Mining-based Fraud Detection Research", School of Business Systems, 2019.

[9] W. Lim, A. Sachan and V. Thing, "Conditional Weighted Transaction Aggregation for Credit Card Fraud Detection", *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pp. 3-16, 2014. Available: 10.1007/978-3-662-44952-3_1 [Accessed 15 December 2019].

[10] [A. Thennakoon, C. Bhagyani, S. Premadasa and S. Mihiranga, "Real-time Credit Card Fraud Detection Using Machine Learning", *Research Gate*, 2019. Available: https://www.researchgate.net/publication/334761474_Real-time_Credit_Card_Fraud_Detection_Using_Machine_Learning. [Accessed 15 December 2019].

[11] F. Carcillo, Y. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé and G. Bontempi, "Combining unsupervised and supervised learning in credit card fraud detection", *Information Sciences*, 2019. Available: https://www.sciencedirect.com/science/article/pii/S0020025519304451. [Accessed 15 December 2019].

[12] N. Carneiro, G. Figueira and M. Costa, "A data mining based system for credit-card fraud detection in e-tail", *Research gate*, vol. 95, pp. 91-101, 2017. Available: https://www.researchgate.net/publication/312255358_A_data_mining_based_system_for_credit-card_fraud_detection_in_e-tail. [Accessed 15 December 2019].

[13] R. Gupta and C. Pathak, "A Machine Learning Framework for Predicting Purchase by Online Customers based on Dynamic Pricing", *Procedia Computer Science*, vol. 36, 2014. Available: https://www.researchgate.net/publication/275541641_A_Machine_Learning_Framework_for_Predicting_Purchase_by_Online_Customers_based_on_Dynamic_Pricing. [Accessed 15 December 2019].

[14] F. Weber and R. Schütte, "A Domain-Oriented Analysis of the Impact of Machine Learning—The Case of Retailing", *Big Data and Cognitive Computing*, vol. 3, no. 1, p. 11, 2019. Available: 10.3390/bdcc3010011.

[15] Kronberger, G. and Affenzeller, M. (2011). Market Basket Analysis of Retail Data: Supervised Learning Approach. *Research Gate*. [online] Available at: https://www.researchgate.net/publication/221431835_Market_Basket Analys is of Retail_Data_Supervised_Learning_Approach [Accessed 15 Dec. 2019].

[16] R. Liu, Y. Lee and H. Mu, "Customer Classification and Market Basket Analysis Using K-Means Clustering and Association Rules: Evidence from Distribution Big Data of Korean Retailing Company", *Research Gate*, 2019. Available: https://www.researchgate.net/publication/330506538_Customer_Classification_and_Market_Basket_Analysis_Using_K-Means_Clustering_and_Association_Rules_Evidence_from_Distribution_Big_Data_of_Korean_Retailing_Company. [Accessed 15 December 2019].

[17] N. Chawla, K. Bowyer, L. Hall and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002. Available: https://www.researchgate.net/publication/220543125_SMOTE_Synthetic_Minority_Over-sampling_Technique. [Accessed 15 December 2019].

[18] C. Tantithamthavorn, A. Hassan and K. Matsumoto, "The Impact of Class Rebalancing Techniques on the Performance and Interpretation of Defect Prediction Models", *IEEE Transactions on Software Engineering*, pp. 7-7, 2018. Available: https://arxiv.org/pdf/1801.10269.pdf. [Accessed 15 December 2019].

[19] S. Mishra et al., "Principal Component Analysis", 2017. Available: https://www.researchgate.net/publication/316652806_Principal_Component_Analysis. [Accessed 17 December 2019].

[20] Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview", in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, *Advances in Knowledge Discovery and Data Mining,* AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34.