PREDICTING YEAR OVER YEAR GROWTH IN INVESTMENT BANKING USING REGRESSION MODELLING

[1]A.Mahesh, [2]Richa, [3]Shubham Tiwari, [4]B.Divya, [5]L. Kiran Kumar Reddy
[1,2,3,4]UG Student, [5]Assistant Professor, [1,2,3,4,5]Dept. of Computer science Engineering, Visvesvaraya College of Engineering and Technology, Mangalpalle, Telangana, India

ABSTRACT
Regression modelling helps organisations use there data to make better decisions.This is done by reliable, data driven logical conclusions about the current and future events and can be achieved by using machine learning techniques to make predictions. Therefore Regression modelling enables the organisations customer focused finding their business issues proactively in realtime and addressing them at right time to get best outcomes.
Application of Regression Modelling Solutions in banking industry include, Segmentation, Application, Fraud Detection, Customer Retention

INTRODUCTION

DOMAIN DESCRIPTION

The science of teaching computers to behave without explicit programming is known as machine learning.
We now have self-driving cars, useful speech recognition, efficient web search, and a much better understanding of the human genome thanks to machine learning over the last ten years. These days, machine learning is probably something you use hundreds of times a day without even realizing it.
It is also regarded by many experts as the most effective path towards human-level AI.The science of teaching computers to behave without explicit programming is known as machine learning. We now have self-driving cars, useful speech recognition, efficient web search, and a much better understanding of the human genome thanks to machine learning over the last ten years. These days, machine learning is probably something you use hundreds of times a day without even realizing it.
It is also regarded by many experts as the most effective path towards AI on par with humans.

Fig 1.1 Machine Learning

Types of Machine Learning:
1. Supervised Learning
2. Unsupervised Learning
3. Reinforcement Learning

1.Supervised Learning:
Supervised learning algorithms are trained using labeled examples, such as an input where the desired output is known as **Supervised learning** algorithms are trained using labeled examples, such as an input where the desired output is known.
2.Unsupervised Learning:
Unsupervised learning is used against data that has no historical labels. The system is not told the "right answer." The algorithm must figure out what is being shown.
3.Reinforcement Learning;
Reinforcement learning is often used for robotics, gaming and navigation. With reinforcement learning, the algorithm discovers through trial and error which actions yield the greatest rewards

About Work

The interactive process of creating mutually beneficial relationships among stakeholders is known as bank direct marketing. A comprehensive understanding of customer traits and behaviors is necessary for efficient multichannel communication. Increasing the response rates of direct promotion campaigns is the aim of bank direct marketing, aside from profit growth, which may boost customer loyalty and positive responses.
Investigations into the available datasets for bank direct marketing analysis have been ongoing. The analysis's goal is to identify target consumer groups that have an interest in particular goods.
Investigations into the available datasets for bank direct marketing analysis have been ongoing. The analysis's goal is to identify target consumer groups that have an interest in particular goods.
Resampling techniques must be used to handle imbalanced datasets. In addition to counteracting the negative effects of imbalance, undersampling and oversampling techniques also improve the prediction accuracy of some well-known machine learning classification algorithms.

Objective

We obtained information from the Reserve Bank of India in order to forecast growth from year to year. In order to forecast growth from year to year, the following factors have been considered: total non-SLR investments, adjusted non-food bank credit, shares, bonds/debentures, fortnight ended, non-food bank credit, investments in commercial paper, and total non-SLR investments.
The central bank's (RBI) main goals are to oversee and carry out financial sector initiatives for commercial banks, financial institutions, and non-banking financial companies (NBFCs). Several important projects include: Changing the way bank inspections are conducted bolstering the function of statutory auditors in the banking industry.
The data set has a lot of imbalance.To improve prediction accuracy, some oversampling techniques are used as a preprocessing step. A regression model is utilized.

Theoretical Background

The work in [6] examined the theoretical underpinnings of marketing analytics, a broad field that emerged from computer science, operations research, marketing, and statistics. They claimed that one of the difficulties with direct marketing analysis is forecasting customer behavior. Additionally, they talked about big data visualization techniques for the marketing sector, including latent Dirichlet allocation, multidimensional scaling, correspondence analysis, and customer relationship management (CRM). They discussed the general trade-off between geographic visualization's common practices and art, as well as how it relates to retail location analysis. They also went into more detail about discriminant analysis as a marketing prediction method. Techniques like ensemble learning, feature reduction, and extraction are all part of discriminant analysis. These methods address issues with customer loyalty, lifetime value, purchase behavior, review ratings, sales, profit, and brand visibility.

EXISTING SYSTEMS

Data from the bank is gathered and examined using the current bank parameters in order to forecast the growth of the investment bank year over year.

Since it's a manual task, it typically takes a while to complete. Large organizations' processes are also difficult to analyze, which makes it difficult to update the data.

Furthermore, additional paperwork will be needed. Making the most of a given dataset can be challenging because it necessitates a thorough analysis of all of its attributes and their values. However, this can be resolved with ease by utilizing machine learning algorithms, which allow us to forecast future events by examining historical data sets that users have supplied to a system.

PROPOSED SYSTEMS

In this system, we give the specifics of the bank data set so that a machine learning algorithm can be used to analyze future results. The machine learning repository at UCI provided the data set. The purpose of data preprocessing is to clean raw data. The technique of regression modeling is utilized. As a supervised machine learning technique for continuous value prediction, regression modeling is an effective method.As a result, it can forecast growth in investment banking year over year based on user inputs into a system.

This system makes use of the NumPy, Pandas, Scikit-learn, and Matplotlib stacks.

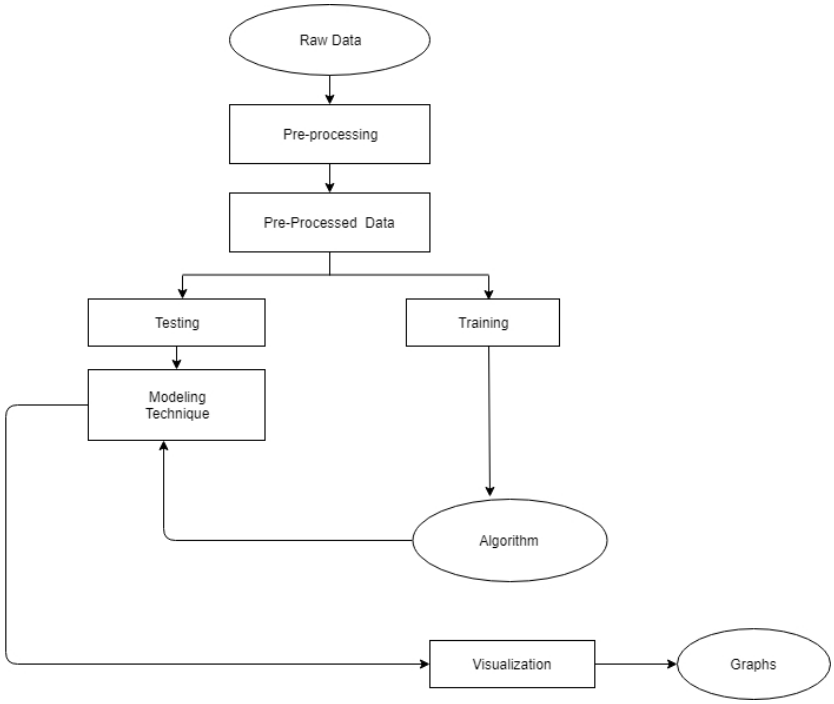
As a result, the system analyzes and predicts the dataset, producing performance models that are visualized as bar and pie charts.

FEASIBILITY STUDY

As the name implies, a feasibility analysis is used to determine the viability of anidea, such as ensuring a project is legally and technically feasible as well aseconomically justifiable. It tells us whether a project is worth the investment—insome cases, a project may not be doable. There can be many reasons for this,including requiring too many resources, which not only prevents those resources from performing other tasks but also may cost more than an organization would earnback by taking on a project that isn’t profitable.

BLOCK DIAGRAM

The block diagram is typically used for a higher level, less detailed description aimed more at understanding the overall concepts and less at understanding the details of implementation.



DATA FLOW DIAGRAMS:

Data flow diagram (DFD) is a graphical representation of "flow" of data through an information system, modelling its process concepts. Often they are a preliminary step used to create an overview of the system which can later be elaborated. DFD's can also be used for the visualization of data processing (structured design).

A DFD shows what kinds of information will be input to and output from the system, where the data will come from and go to, and where the data will be stored. It doesn't show information about timing of processes, or information about whether processes will operate in sequence or parallel. A DFD is also called as "bubble chart".

IMPLEMENTATION

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective.

The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

The project is implemented by accessing simultaneously from more than one system and more than one window in one system. The application is implemented in the Internet Information Services 5.0 web server under the Windows XP and accessed from various clients.

Technologies Used

What is Python?

"Guido van Rossum" created Python, a high-level interpreter language for general-purpose programming, which was first made available in 1991. Python's design philosophy places a strong emphasis on readability of code, and its syntax enables programmers to convey ideas in fewer lines of code—notably with ample whitespace. It offers structures that make programming understandable at both small and large scales.

Python has automatic memory management and a dynamic typing system. It has a sizable and extensive standard library and supports a variety of programming paradigms, such as imperative, functional, procedural, and object-oriented programming.

TESTING

It is the process of running a program with the goal of identifying errors and testing the functionality. A test case that has a high likelihood of discovering an as at undiscovered error is considered good. A test that finds an as-yet-undiscovered error is considered successful. One of two reasons is typically involved in software testing: Defect Detection

Reliability estimation

Results

We have Machine Learning classifier models to evaluate. In that we have used Linear Regression Model to predict the growth year by year in investment banking.

import modules

```
In [2]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
        from sklearn.preprocessing import StandardScaler #StandardScaler function of sklearn is used to scale down values of dataframe b
        from sklearn.model_selection import train_test_split
        from sklearn.metrics import roc_auc_score
        from sklearn.ensemble import AdaBoostRegressor #import adaboostregressor from sklearn.ensemble class
        from sklearn.metrics import mean_squared_error  #retrieving the training(x_train,y_train) accuracy score of the AdaBoostRegressor
        from sklearn.ensemble import RandomForestRegressor #import Random forest regressor from sklearn.ensemble class
        #from tpot import TPOTRegressor
```

Data Loading

```
In [3]: #using pd.read_csv method from pandas to load the csv sheet into a dataframe named df
        df = pd.read_excel("C:\\Users\\SWAYAM\\Downloads\\Project\\Flow of Financial Resources from Scheduled Commercial Banks to the Com
```

```
In [4]: df.shape
Out[4]: (66, 9)
```

```
In [5]: df
```

```
In [5]: df
Out[5]:
```

	Unnamed: 0	Fortnight Ended	Non-food Bank Credit	Investments in Commercial Paper	Investments in shares	Investments in Bonds/Debentures	Total Non-SLR Investments	Adjusted Non-food Bank Credit	Y-o-Y Growth in (7) (%)
0	NaN	1	2.000000e+00	3.0000	4.0000	5.0000	6 = (3 to 5)	7 = (2 + 6)	8.000000
1	NaN	2018-09-27 00:00:00	9.711225e+06	93828.7379	83326.5350	543412.0020	721067	1.04332e+07	8.474251
2	NaN	2018-09-13 00:00:00	9.646922e+06	94406.4320	83370.0320	536088.0210	716444	1.03634e+07	8.619415
3	NaN	2018-08-30 00:00:00	9.617761e+06	95878.8334	83847.1103	535545.3520	715271	1.03378e+07	9.709062
4	NaN	2018-08-16 00:00:00	9.610342e+06	98241.7326	83503.1477	537506.1614	719351	1.03385e+07	11.037356
...									
61	NaN	2017-06-09 00:00:00	7.573360e+06	125872.1223	74300.1649	342207.7803	542380	8.11574e+06	7.035484
62	NaN	2017-05-26 00:00:00	7.508757e+06	114847.2761	72732.1614	341836.9361	529415	8.03817e+06	6.106126
63	NaN	2017-05-12 00:00:00	7.532945e+06	161517.3561	72740.3156	346105.0417	562451	8.1154e+06	7.199091
64	NaN	2017-04-28 00:00:00	7.526710e+06	111732.2515	73479.7016	342035.0500	527247	8.05396e+06	6.256098
65	NaN	2017-04-14 00:00:00	7.553426e+06	110751.6766	71236.0045	346802.1672	530790	8.08422e+06	6.607951

56 rows × 9 columns

Data Preprocessing

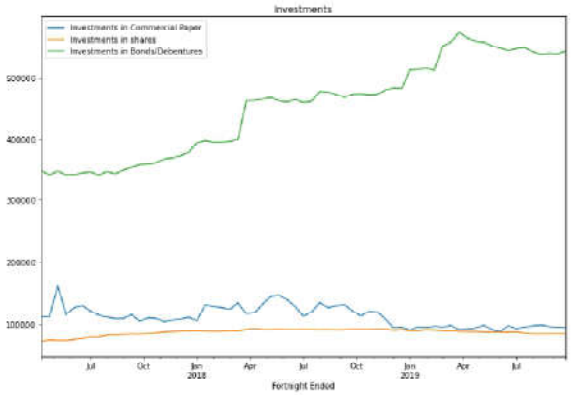
```
In [6]: df.drop(0,axis=0,inplace=True) #dropping the row with index 0 using drop method of pandas Dataframe
```

```
In [7]: df.drop('Unnamed: 0',axis=1,inplace=True) #dropping the unnamed column as it has only NaN values along vertical axis
```

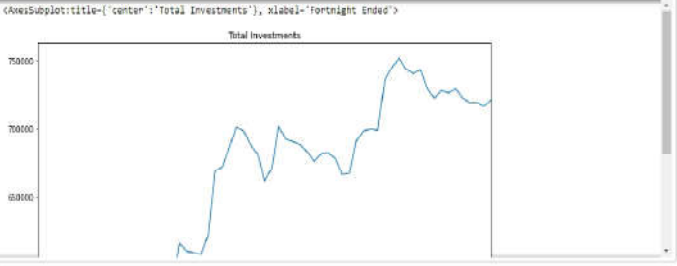
```
In [8]: df.head(10)  #displaying the head of dataframe(head displays first 5 rows)
Out[8]:
```

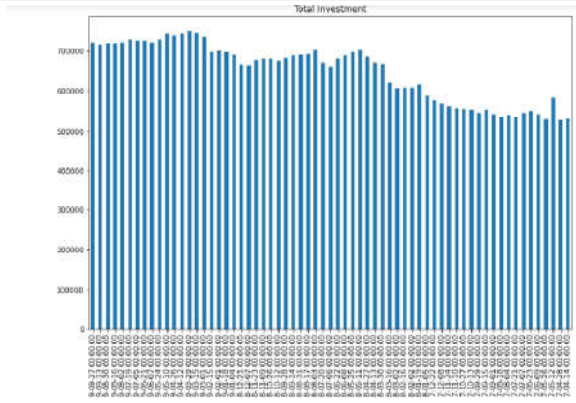
	Fortnight Ended	Non-food Bank Credit	Investments in Commercial Paper	Investments in shares	Investments in Bonds/Debentures	Total Non-SLR Investments	Adjusted Non-food Bank Credit	Y-o-Y Growth in (7) (%)
1	2019-09-27 00:00:00	9.711225e+06	93828.7379	83626.5360	543412.0029	721067	1.04323e+07	8.474251
2	2019-09-13 00:00:00	9.646922e+06	94486.4220	83676.0929	538088.0219	716444	1.00834e+07	9.619415
3	2019-08-30 00:00:00	9.617761e+06	95878.8334	83847.1103	539545.3529	719271	1.0337e+07	9.709062
4	2019-08-16 00:00:00	9.619342e+06	96241.7826	83603.1477	537506.1614	719351	1.03387e+07	11.037356
5	2019-08-02 00:00:00	9.666265e+06	96892.3612	83606.2974	541910.6789	722409	1.03887e+07	11.408090
6	2019-07-19 00:00:00	9.582377e+06	94921.0782	84616.2342	548229.8709	729960	1.03213e+07	11.032259
7	2019-07-05 00:00:00	9.626666e+06	92114.1328	86773.6693	547319.4120	726207	1.03529e+07	11.818919
8	2019-06-21 00:00:00	9.577349e+06	97238.6098	86444.6681	544187.3584	727872	1.03052e+07	11.607647
9	2019-06-07 00:00:00	9.574300e+06	97263.2560	86574.3462	548389.7587	722227	1.02965e+07	11.720183
10	2019-05-24 00:00:00	9.555823e+06	90752.7139	86804.7128	551955.7108	729513	1.02683e+07	12.003225

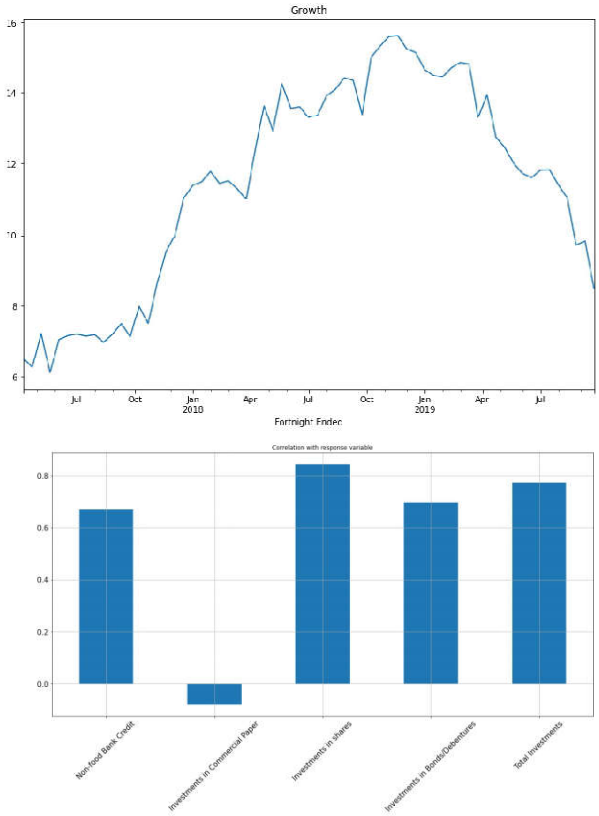
```
Out[15]: <matplotlib.legend.Legend at 0x2817521e2e8>
```




```
In [17]: df['Total Investments'].plot(figsize=(12,8),title='Total Investments')
Out[17]: <AxesSubplot:title={'center':'Total Investments'}, xlabel='Fortnight Ended'>
```







Hence by using Regression model we have predicted the growth year by year in investment banking.

CONCLUSION

This research aimed to provide a visualization mechanismfor simple classification tasks. Experiments were conducted on an imbalanced data set of a RBI. Hence by using Regression model we have predicted the growth year by year in investment banking. Random Forest Regressor is used, which gives the accuracy between 65%-99%.

FUTURE ENHANCEMENTS

In future we can use deep learning, machine learning is the super set of deep learning which is considered one of the useful methods for predicting growth by growth in investment banking. In future by advancing the technology we can predict the growth with much more accuracy.

References

1. Statistics-YouTube. http://www.youtube.com/yt/press/statistics.html, 2013.
2. J. Friedman, T. Hastie, and R. Tibshirani. The elements of statistical learning: Data Mining, Inference, and Prediction, Second Edition. Springer Series in Statistics, 2009.
3. G. Szabo and B. Huberman. Predicting the popularity of online content. Communic. Of ACM, 53(8), 2010.
4. A Machine Learning Model for Stock Market Prediction Article by Osman Hegazy and Mustafa Abdul Salam.
5. The Unified Modeling Language User Guide, Low Price Edition Grady Booch, James Rumbaugh, Ivar Jacob**,** ISBN: 81-7808-769-5, 91, 1997.

6. The Elements of UMLTM 2.0, Scott Ambler-Cambridge University Press Newyork@2005,ISBN: 978-0-07-52616-7-82, 2005.
7. Software Testing, CemKaner, James beach, BretPettiehord, ISBN: 978-0-471-120- 940,2001.
8. A Practitioner's Approach Roger S. Pressman, Software Engineering, 3rd Edition, ISBN: 978-007-126782-3,482, 2014.
9. Black_Box Testing: Techniques for Functional Testing of Software and Systems Boris Beizer - Wiley Publications, ISBN: 978-0-471-120-940, 1995.