**Capstone Project** **Document**

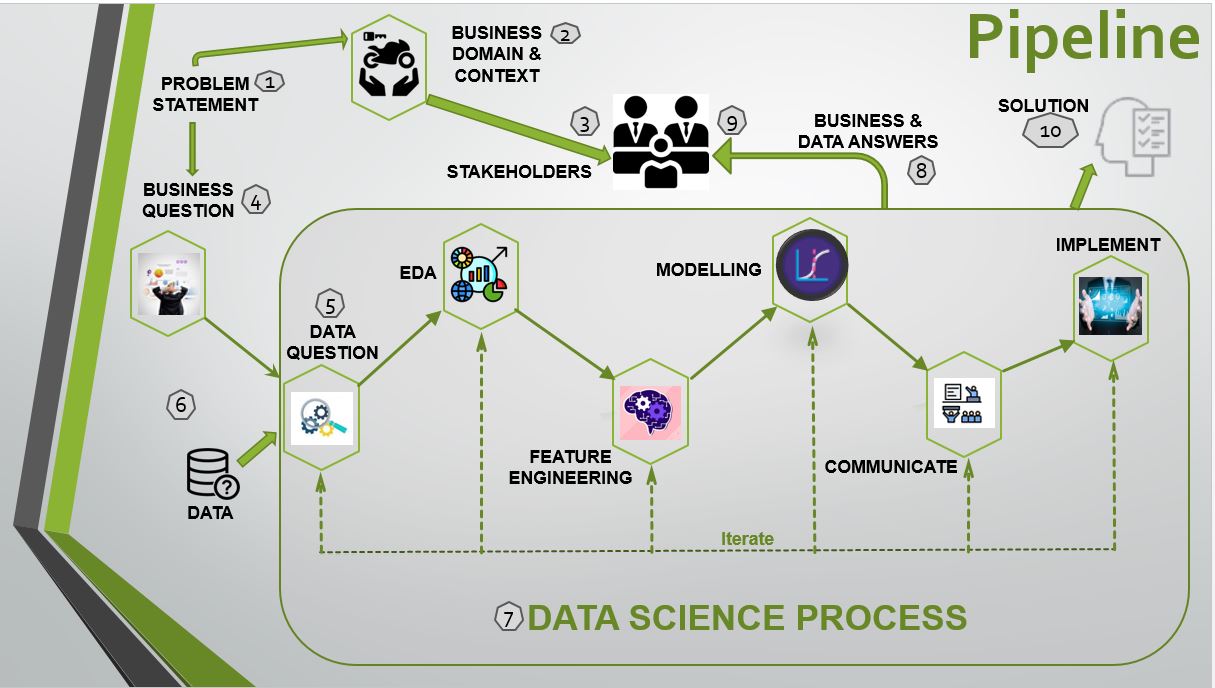
DSIA – Data Science and AI Course

Part-time – 2021-05-29

**MONICA ESTRADA**

# Process overview

The following diagram shows the overall end-to-end process for defining, designing and delivering the Capstone project.



# Problem statement

* **What is the problem or the opportunity that the project is investigating?**

Assess the risk of a customer defaulting on a motorbike loan.

* **Why is this problem valuable to address?**

India has seen an increase in two-wheelers loan defaults since 2007 and in 2020 the financing over them decreased by 50%. Considering that the large part of the Indian market is made up low income group (who can only afford the low-cost entry level of motorcycles), make them vulnerable in extreme situations like a lockdown. On the other hand the two-wheeler loan market in India is projected to grow significantly.

Even though this situation represents a problem at the same time offers an opportunity since financiers could develop robust and trustworthy tools to assess loan applications with more accuracy and speed to cope up with the predicted demand.

Motorbikes dealers can get bankers to finance their inventory and also make more sales if the loan risk is decreased.

Loan customers can have the result of their applications fast.

Credit analysts are gone, it’s the machine learning age! Machine learning algorithms have a lot to offer to the world of credit risk assessment due to their unparalleled predictive power and speed.

* **What is the current state?**

Two-wheeler loan defaults in India have been on the rise (much more after the Covid-19 struck) (3). Consequently, financiers have been adopting stricter process to screen customer profiles (2) and motorbikes dealers have been finding difficult to get banks to finance their inventory(3).

* **What is the desired state?**

From the financiers’ perspective, being able to approve loans only to customers that are not at risk on defaulting during the length of the loan, rapidly and accurate.

From motorbikes dealers be able to finance their inventory and also offer loans with low risk.

* **Has this problem been addressed by other research projects? What were the outcomes?**

Developed credit risk modelling have been based on either financial statement analysis, default probability, or machine learning.

Main variables that have been used when assessing low credit score are: previous or existing loans, delayed payment, credit card bills, missing payments, and multiple credit applications

# Industry/ domain

* **What is the industry/ domain?**

The industries domains for the present project are finance and transportation.

* **What is the current state of this industry? What is the overall industry value-chain? What are the key concepts in the industry?**

The two-wheeler loan market in India is projected to grow from an estimated US$7.2 billion in 2020 to US$12.3 billion by 2025. Major players operating in the Indian Two Wheeler Loan Market are categorized into Non-Banking Finance Companies (NBFC) and banks (15).

The Indian Two Wheeler Loan Market can be segmented based on: type, source, percentage of amount sanctioned, type of city, tenure, and region (15).

Worldwide motorcycles market had total revenues of US$69,401m in 2016 (7). On a global scale motorcycles market is concentrated in the Asian-Pacific region countries, an area that accounts for 80% of total annual global sales (14).

Within the overall transportation industry, the motorcycle sector is experiencing significant revenue growth. Motorcycles are designed for distance travel, commuting, and off-road racing. On the global market, the industry is characterized by the slow diversification of demand for new fuel types and by new geographic regions affected (9).

According to the latest market research report, the global motorcycles market is expected to grow of 5.73% during the forecast period from 2020 to 2027. The rise in the expendable income of the middle-class section across the globe has given rise to the growth of the demand for the motorcycle, due to its affordability and suitable modes of transportation. The demand for the motorcycle from the young population is expected to drive the growth of the Motorcycle Market in developing countries such as India, the US, Brazil, China, and Mexico (10).

In the year 2020, the global motorcycle market trend was predicted to have a significant impact due to the outbreak of the Covid-19. The measures from the government and lockdown due to the Covid-19 have immobilized the retail business of motorcycles across the world. The global pandemic has disturbed the supply chain of motorcycles. Furthermost, the public transport have motivated the consumers in the purchasing the motorcycle in 2020-21 (10).

After China, India is the second-largest market of motorcycles due to some key factors including fuel-efficiency, inexpensiveness, and the low displacements of the bikes. This country has witnessed tremendous growth in the demand of the motorcycle from the rural areas of the country due to the construction of the roads and due to the development of the infrastructure. On the basis of the capacity of the engines, the 51-125cc segments hold the maximum motorcycle market shares, which is about 50% and the 126-250cc segments hold about 52% of the market shares (10).

# Stakeholders

* **Who are the stakeholders? (be as specific as possible)**

Financial institutions

Motorbikes dealers

* **Why do they care about this problem?**

Loans make up the biggest risk for any bank and because of that, the banking industry has been focusing more attention than ever on credit management. To ensure their healthy financial status, banks must balance their risks daily by having credit risk (11).

Having proper credit risk management is important for lenders since it can help lowering the capital that is locked with the debtors and therefore manage cash flow more efficient, reducing the possibility of getting into bad debts, improving profits and enhancing customer management.

* **What are the stakeholders’ expectations?**

As the banking rules become more complex and the consequences of noncompliance more severe, banking will become more automated. Some banks have started to experiment with machine learning models in collections or credit-card-fraud detection and credit decisions (7).

# Business question

* **What is the main business question that needs to be answered?**

How to assess the risk of a loan default fast and accurate?

* **What is the business value of answering this question**

Loan default risk is simply known as the possibility of a loss for a lender due to a borrower’s failure to repay a loan. Credit analysts are typically responsible for assessing this risk by thoroughly analysing a borrower’s capability to repay a loan. Machine learning algorithms have a lot to offer to the world of credit risk assessment due to their unparalleled predictive power and speed.

* **What is the business value of answering this question? (quantify value and make necessary assumptions)**

If Indian total value motor loan market 2020 is US$7.2b & increase ratio on bad loans ratio 3.5% for 2021, decreasing only 20% of that ratio would represent a decrease on bad loans on near US$504 million.

* **What is the required accuracy? What are the implications of false positives or false negatives?**

From the perspective of a financial services organization, there are two kinds of adverse events that can occur: False positives are rejections of applicants that would not have defaulted, which reduce the organization’s profits. False negatives are approvals of applicants that do ultimately default, which increase the organization’s default risk. The costs of these two kinds of adverse events are not equal.

F1-score (the trade-off between recall and precision) is a very good metric for predicting loan default.

# Data question

* **What is the data question that needs to be answered?**

What model could be used to predict risk of defaulting on a personal loan and what are the metrics?

* **What is the data required to answer the question?**

What are the main variables to look up when assessing a personal loan application?

# Data

* **Where was the data sourced?**

The data, authored by Sayantan Jana in July 2020, was retrieved from Kaggle and it was given to participants in a tech competition.

* **What is the volume and attributes of the data?**

Original dataset has 119,528 observations, 30 features and one target variable (1= Defaulters and 0= Non-Defaulters).

1. V1: Customer ID = INDEX

2. V2: If a customer has bounced in first EMI (1: Bounced, 0: Not bounced)

3. V3: Number of times bounced in recent 12 months

4. V4: Maximum MOB (Month of business with TVS Credit)

5. V5: Number of times bounced while repaying the loan

6. V6: EMI (Equated Monthly Instalments)

7. V7: Loan Amount

8. V8: Tenure = the period or duration for which the loan amount is sanctioned

9. V9: Dealer codes from where customer has purchased the two wheeler

10. V10: Product code of two wheeler (MC: Motorcycle, MO: Moped, SC: Scooter)

11. V11: No of advance EMI paid

12. V12: Rate of interest

13. V13: Gender (Male/Female)

14. V14: Employment type (HOUSEWIFE: housewife, SELF: Self-employed, SAL: Salaried, PENS: Pensioner, STUDENT: Student)

15. V15: Resident type of customer

16. V16: Date of birth

17. V17: Age at which customer has taken the loan

18. V18: Number of loans

19. V19: Number of secured loans: Number of unsecured loans

20. V21: Maximum amount sanctioned in the live loans

21. V22: Number of new loans in last 3 months

22. V23: Total sanctioned amount in the secured Loans which are Live

23. V24: Total sanctioned amount in the unsecured Loans which are Live

24. V25: Maximum amount sanctioned for any two wheeler loan

25. V26: Time since last Personal loan taken (in months)

26. V27: Time since first consumer durables loan ta (in months)

27. V28: Number of times 30 days past due in last 6 months

28. V29: Number of times 60 days past due in last 6 months

29. V30: Number of times 90 days past due in last 3 months

30. V31: Tier; (Customer’s geographical location)

31. V32: Target variable (1: Defaulters / 0: Non-Defaulters)

* **How reliable is the data?**

On Kaggle there is not information about the reliability of the data.

* **What is the quality of the raw data?**

The data, as expected for its nature, is highly unbalanced (since defaulters constitute a rare event).

* **How was this data generated?**

The dataset was created on 29th July 2020 and was given to participants in a tech competition.

* **Is this data available on an ongoing basis?**

No, there won’t be any update.

# Data science process

## Data analysis

* **What data pipeline was to wrangle the raw data?**
* Description (shape, types, info, describe) to have a general understanding;
* Splitting original dataset (highly imbalanced) into defaulters and non-defaulters, subsampling the majority class to make equal in size to the minority one;
* Concatenating both subsets datasets into a new one;
* Dropping variables:
  + Five features with the majority of values missing were removed (a feature that has a high number of empty values is unlikely to be very useful for prediction). These features were: Time since last Personal loan taken (in months), Time since first consumer durables loan taken (in months), Total sanctioned amount in secured Loans which are Live, Total sanctioned amount in unsecured Loans which are Live, and Max amount sanctioned loans;
  + Two other features were also removed since they did not have too much importance (dealer code) and non-meaningful information (tier, which was associated to geographic location but no details were offered only numbers from 1 to 4);
  + Another removed feature was ‘Number of new loans in last 3 months’ since there were only zero values;
  + Birth date was also removed since there was another feature named ‘Customer Age when taken loan’ that was highly correlated and with similar information;
* Profiling (correlations, distributions, understanding relationships and new insights).
* **What are the highlights of the Exploratory Data Analysis (EDA)?**
* Strong positive correlation (+0.71) between number of times bounced while repaying loan with number of times bounced in recent 12 months;
* Moderate positive correlation (+0.65) between loan amount and EMI (EMI: Equated Monthly Instalments);
* Moderate positive correlation (+0.62) between number on unsecured loans and number of loans;
* Moderate positive correlation (+0.59) between maximum amount sanctioned for any two wheeler loan and loan amount;
* Moderate negative correlation (-0.61) between product code two wheeler MO and loan amount;
* There was not a particular variable to have a significant correlation with the Target (defaulter or non-defaulter). The highest value of +0.32 was with number times 30 days past due in last 6 months;
* Many features have distributions with many outliers and therefore can cause a significantly impact on the mean and standard deviation and are more difficulty for a model to classify them correctly. Furthermore, some of the features select by random forest have this kind of distribution: EMI (Equated Monthly Instalments), loan amount, number of times 30 days past due in last 6 months, and number of times bounced while repaying the loan.
* **Is the pipeline reusable?**

Yes, the pipeline is reproducible.

## Modelling

* **What are the main features used?**

Random Forest Classification Feature Importance was used to select the features to be included in the models. This embedded method was selected since it determines the final importance of each variable, is highly accurate, generalize better and is interpretable.

* **What feature engineering techniques are used?**
* Imputation to replace missing data (numerical and categorical) using TransformerMixin which used the most frequent value for categorical variables and the mean for the numerical variables;
* One hot encoding to transform some categorical features into numerical;
* Scaling using StandardScaler for input variables to normalise the data.
* **What are the models used?**

**MACHINE LEARNING METHODS**

* + They are able to learn from patterns of normal behaviour;
  + Fast to adapt to changes in the normal behaviour and can identify patterns of defaulters; therefore can identify risky potential customers even before they have been granted a loan;
  + Five models selected: KNN, Super Vector Machine, Decision Tree, Random Forest, and Logistic Regression, with the 8 feature importance variables from Random Forest Classification Feature;
  + Hyper parameter optimization for the best two models (Random Forest and Logistic Regression).

**ENSEMBLE LEARNING METHODS**

* + Stacking with the five models (using default hyper-parameters);
  + Boosting with the best two models (with optimized hyper-parameters). The AdaBoost model can increase the weights of misclassified points on every iteration and therefore might put high weights on the outliers.
  + Hard Voting (for the 2 best models with and without hyper-parameter optimization and boosting).
* **How long does it take to train your model?**

A few minutes, with the clean data set.

Optimizing the hyper-parameters of the two best models can take up to three to four hours.

* **What are the tools used?**

Jupyter notebook and Jupyter Lab

* **What are the model performance metrics?**

A common way to evaluate the performance of a model with binary responses is to use the confusion matrix. The observed cases of default are defined as positives and non-default as negatives. The possible outcomes are then true positives (TP) if defaulted customers have been predicted to be defaulted by the model. True negatives (TN) if non-default customers have been predicted to be non-default. False positives (FP) if non-default customers have been predicted to be defaulted, and false negatives (FN) if defaulted customers have been predicted to be non-default.

* + Accuracy: the problem with accuracy when applying it for imbalanced data is that it can be high but only predicting the majority class. So specificity is more relevant for this kind of situations
  + Confusion Matrix: A breakdown of predictions into a table showing correct predictions and the types of incorrect predictions made (.
  + Precision: A measure of a classifiers exactness
  + Recall: A measure of a classifiers completeness
  + F1 Score : weighted harmonic average of precision and sensitivity
  + ROC-AUC: relation between true-positive rate and false positive rate
* **Which model was selected?**

Logistic Regression and Random Forest were selected as best models. Hyper-parameter optimization and boosting were also implemented for both models in order to assess if there was a significant improvement in their metrics.

From the hard voting the number one model is random forest with optimized parameters (mean accuracy of 0.735 and standard deviation of 0.042); however the boosted logistic regression with default hyper-parameters had a very close result (0.73 as accuracy mean and 0.042 standard deviation). Considering that random forest is a more complex model using more computational power, logistic regression is the best option since it is simpler, easy to interpret and uses less computational power. But, it is also important to highlight that the logistic regression is more impacted by outliers than the random forest model.

Looking at the other metrics, particularly to F1-score, both models have very close results as well: 74.23% for logistic and 73.42% for random forest.

## Outcomes

* **What are the main findings and conclusions of the data science process?**
  + From the Random Forest Classification Feature Importance eight features (out of 21) %. They are listed in the table below (in order of importance):

|  |  |  |
| --- | --- | --- |
| **Feature** | **Name** | **Score** |
| 12 | Number of times 30 days past due in last 6 months | 0.28208 |
| 11 | Maximum amount sanctioned for any Two wheeler loan | 0.14954 |
| 9 | Number of loans | 0.08803 |
| 3 | EMI (Equated Monthly Instalments) | 0.07263 |
| 8 | Customer age when taken loan | 0.07243 |
| 4 | Loan Amount | 0.06834 |
| 7 | Interest rate | 0.06009 |
| 1 | No times bounced 12 months | 0.04237 |
|  | **TOTAL** | **0.83551** |

* + Even if SVM had slightly better results than logistic regression, the second one was selected as a second option since SVM is a more complex model that may suffer from overfitting.
  + The voting ensemble did not provide a significant better performance that any single model used in the ensemble. Therefore, any model that performs better should be used instead. In this case random forest with hyper-parameters or AdaBoost logistic regression with default hyper-parameters.
  + The voting ensemble is not guaranteed to provide better performance than any single model used in the ensemble. If any given model used in the ensemble performs better than the voting ensemble, that model should probably be used instead of the voting ensemble.
  + The table below presents the evaluation metrics for both models.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **METRICS** | **Logistic Regression** | **Random Forest** |
| Metrics **Default** Hyper-parameters | Cross-validation | 74.10% | 75.39% |
| Accuracy | 72.75% | 72.66% |
| Recall | 62.52% | 69.98% |
| Precision | 78.61% | 73.94% |
| F1-Score | 69.65% | 71.91% |
| Metrics **Optimized** Hyper-parameters | Cross-validation | 74.99% | 77.52% |
| Accuracy | 74.38% | 74.67% |
| Recall | 73.81% | 69.98% |
| Precision | 74.66% | 77.22% |
| F1-Score | 74.23% | 73.42% |

## Implementation

* **What are the considerations for implementing the model in production?**

The model still needs more improvement and more data before its implementation.

# **Data answer**

* Was the data question answered satisfactorily?

Yes, two models were identify that are able to predict the risk of defaulting on a personal loan: Random Forest and Logistic Regression.

Boosted Logistic Regression was selected as the best option. Some of the pros and cons are listed below:

* **Pros:** stable, simpler, easier to implement and interpret, efficient to train, uses less computational power. Very popular and is used when target is categorical. It is easy to convert the model result to a specific strategy and deploy.
* **Cons:** easy to under-fit, high demand for data, not to good at dealing with unbalanced data, high-dimension feature set, and categorical features.
* **Boosting:** makes the model more robust and stable, improves accuracy, but reduce interpretability.

There are eight main variables to look up when assessing a personal loan application (loot at the table above).

* **What is the confidence level in the data answer?**

Both models have similar evaluation metrics. Accuracy around 74% and F1-score (trade-off between recall and precision) around 73-74%.

# Business answer

* **Was the business question answered satisfactorily? What is the confidence level in the business answer?**

Yes, machine learning models can offer a solution when assessing the risk of a loan default fast. However the accuracy and F1-score are not as high as desirable; further improvement is recommended.

Classification of imbalanced data and outliers’ management are a challenge for machine learning.

# Response to stakeholders

* **What are the overall messages and recommendations to the stakeholders?**

Machine learning algorithms have a lot to offer to the world of credit risk assessment due to their unparalleled predictive power and speed.

# End-to-end solution

* **What is the overall end-to-end solution to use the model developed in the project?**

Even if the Boosted Logistic Regression model with optimized hyper-parameters is the best option, further steps can be underrating to improve the accuracy, and particularly the precision and recall. Some of these steps could be:

* + Model performance could be improved with more rich data (especially from the defaulters, minority class).
  + Need of more domain knowledge to deal with outliers. Data outliers can spoil and mislead the training process resulting in longer training times, less accurate models and ultimately poorer results.
  + Domain knowledge would help in assessing which outliers should be kept, which removed and/or which to change for another variable.
  + Use penalized models to impose an additional cost on the model for making classification mistakes on the minority class during training and see how performance improves.

# References

* **Where are the data and code used in the project?**

Available at: github.com/MonicaEstrada888

* **What are the resources used in the project? (libraries, algorithms, etc)**

Available at: github.com/MonicaEstrada888

* **Other sources**

**1** Baggonkar, Swaraj, 2020, *This week in Auto: Banks unwilling to lend to two-wheeler buyers over default fears,* Money Control, viewed 6 May 2021, <https://www.moneycontrol.com/news/technology/auto/this-week-in-auto-banks-unwilling-to-lend-to-two-wheeler-buyers-over-default-fears-5947271.html>

**2** Beshr, Sarah, 2020. *A machine learning approach to credit risk assessment.* Towards data science, viewed 16 May 2021, <https://towardsdatascience.com/a-machine-learning-approach-to-credit-risk-assessment-ba8eda1cd11f>

**3** Brownlee, Jason, 2020, *How to develop voting ensemles with python,* Machine Learning Mastery, viewed 16 May 2021, <https://machinelearningmastery.com/voting-ensembles-with-python/>

**4** Coverfox, 2021, *CIBIL Score for Bike Loan*, Coverfox, viewed 6 May 2021, <https://www.coverfox.com/cibil/cibil-score/cibil-score-for-bike-loan/>

**5** CRISIL, 2019, *Ratings, Rating Rationale*, viewed 6 May 2021, <https://www.crisil.com/mnt/winshare/Ratings/RatingList/RatingDocs/WheelsEMI_Private_Limited_April_08_2019_RR.html>

**6** [Dhamija](https://timesofindia.indiatimes.com/toireporter/author-Anshul-Dhamija-479164682.cms), Anshul, 2007, *Two-wheeler loan defaults on the rise*, The Times of India (TOI), viewed 6 May 2021, <https://timesofindia.indiatimes.com/business/india-business/two-wheeler-loan-defaults-on-the-rise/articleshow/2423112.cms>

**7** Harle, Phillipp, et.al, 2015, *The future of bank risk management*. McKinsey & Company, viewed 21 May 2021, <https://www.mckinsey.com/~/media/mckinsey/dotcom/client_service/risk/pdfs/the_future_of_bank_risk_management.ashx>

**8** MarketLine, 2018, *Global Motorcycles Industry Profile & Value Chain Analysis*, viewed 5 May 2021, <https://store.marketline.com/report/vc0199-0403--global-motorcycles-industry-profile-value-chain-analysis/#tab-list-of-tables>

**9** Market Research.com *Motorcycles Market Research Reports & Industry Analysis*, viewed 5 May 2021, <https://www.marketresearch.com/Heavy-Industry-c1595/Transportation-Shipping-c95/Motorcycles-c1650/>

**10** Market Research Future, 2020, *Global Motorcycles Market Research Report: Information By Type (Adventure, Cruiser, Sports, Standard and Touring) and Country - Forecast till 2027*, viewed 5 May 2021, <https://www.marketresearchfuture.com/reports/motorcycles-market-10217>

**11** Mkwawa, Kelvin, 2019. *The importance of credit risk management in banks.* Linkedin, viewed 21 May 2021, <https://www.linkedin.com/pulse/importance-credit-risk-management-banks-kelvin-mkwawa-mba/>

**12** MotorCycle Data, MCD team, 2021, *Global Motorcycles Market lost 9 million sales in* 2020. viewed 6 May 2021, <https://www.motorcyclesdata.com/2020/12/22/world-motorcycles-market/>

**13** Sayantan, Jana, (2020, August), TVS\_Loan\_Default, Dataset for predicting defaulters, Version 1, Retrieved Febraury 26, 2021 from <https://www.kaggle.com/sjleshrac/tvs-loan-default/version/1>

**14** Yongpisanphob Wanna, 2019, *Industry Outlook 2019-2021: Motorcycle*, viewed 5 May 2021, <https://www.krungsri.com/en/research/industry/industry-outlook/Hi-tech-Industries/Motorcycles/IO/io-motorcycles-20>

**15** Yahoo!finance, 2020, *Market Trends & Developments in India's Two Wheeler Loan Market 2020-2025*, viewed 15 May 2021, <https://finance.yahoo.com/news/market-trends-developments-indias-two-114500571.html>