Prediction Of Future Sales

Harshini S[1], Ishwarya Lakshmi S[2], Janani M[3], Dr. G. Fathima M.E, Ph.D[4]
[1, 2, 3,4] Dept of Computer Science And Engineering
[1, 2, 3] Adhiyamaan College of Engineering
[4]Head of the Department, Adhiyamaan College of Engineering

Abstract- *Techniques for making future predictions based upon the present and past data, has always been an area with direct application to various real life problems. The approach shown in this paper is a systematic and precise model building to be used in computing and predicting future sales projection of a product in the market . Forecasting is an important factor, in any organization, to plan and deliver products. Real market data provided by a retailer have been used here to cover numerous categories of data from several types of stores. The problem statement is provided by www.kaggle.com , which also serves as an ongoing competition on the Kaggle platform. In this project we are working with achallenging time-series dataset consisting of daily sales data, provided by the largest Russian software firms- 1C Company. We are predicting total sales for every product and store in the next month by using Random Forest algorithm which is benchmarked against an ensemble of regression trees. The proposed method significantly improves the accuracy of the forecast in many diverse categories andgeographical locations, yielding significant and operative benefits for the manufacturers and the retailers.*

I. INTRODUCTION

Accurate sales forecasting is of utmost value to any organization, be it a small and medium enterprise or a Fortune 500 company. It provides an accurate estimation of company's top-line growth and can be utilized to prepare plans for near-term demand and supply actions. Also, it serves as a guidance to devise the bottom-line plans that keep up with the overall organizational goals for financial prudence. Finally, strategic decisions such as identification of key areas for short-term investment can also be made, taking cues from these forecasts. Sales forecasting can have a crucial impact on the success and performance of companies. Inaccurate forecasts presumably lead to stock-outs or over stock inventories which result in losses for the companies. In particular, within the retail and consumer-oriented industries, such as the electronic market or the fashion industry, accurate forecasts are essential. Companies face several challenges regarding accurate forecasts. For instance, they have to place their production plans before exact knowledge about future demands is available. This is required due to the fact that most production plants are located in Asian countries and therefore the time-to-market is longer than the selling period of products. Therefore, accurate forecasts are crucial because the production of successful products is hardly possible. In addition, other factors, such as changing weather conditions, holidays, public events as well as the general economic situation, can have an impact on future demands. Moreover, fashion items are replaced mostly for the following season; therefore, there is a huge lack of historical sales data. Summing up, due to short life cycles, high variability in products and demand uncertainties, fashion companies often face high challenges with regard to precise forecasts.

Introducing new products to the market, companies have to deal with a lack of historical sales data. Motivated by this problem, researchers have proposed different forecasting models for this special scenario. These works will also be presented. Other review papers on sales forecasting do not consider user-generated content as a potential factor within this research field. First, determine the average length in days of your sales process. This figure is also known as timetopurchase or sales velocity. Add the total number of days it took to close all of the past year's deals and divide by the number of deals. Then, calculate the probability of new deals closing in a certain period of time as a percentage of the average sales cycle length. With this method, the biases of individual reps are less of a factor than with the deal stage model. Also, with this technique, you can fine-tune the probabilities for different lead types. But this technique requires you to know and record how and when prospects enter your pipeline, which can be time intensive.

Interesting and difficult sales forecasting problems are pretty common. We are very fortunate that a similar problem statement titled "Predict Future Sales" is provided by Kaggle. As part of the problem statement, we need to work over a challenging time-series dataset consisting of daily sales data. A time series is a sequence of historical measurements of one or more observable variables at equal time intervals. Time series can be studied for several purposes such as predicting future based on past knowledge, or understanding the latent variables behind the generation of measured values, or for simply providing a concise description of the salient features of the series.

II. OBJECTIVE

The objective of this project is to forecast the total sales of every product and store combination for the next month, given the past data. The sales forecast aims to predict future sales and is used as the basis of planning time and resources. The good forecast should have several objectives, all directly at identifying what we will sell, when we will sell it and to whom.

We predict how many units we are going to sell each month using the Random Forest method in order to understand the dataset better. Then, we figure out what the highest and lowest price is going to be for each unit and the total sales we plan on making each month.

III. LITERATURE SURVEY

1. Mining the past to determine the future market: Sales forecasting using TSDM frameworkAuthors: AngliqueD.Lacasandile, Jamin d. Niguidula, Jonathan M. CaballeroDate of Publication: November 5-8,2017Methodology: Time Series Data Mining

Summary:This study aims to determine patterns of sales behavior to improve service quality. The emphasis is on the aspect discerning pattern from the mass data to have meaningful and useful information that can be used as a basis for decision making particularly in predicting sales trend throughout the year. The paper presents a review of trends in four branches of COPYTRADE, using data retrieved from January 2014 to December 2015.

2. Neural network-based model for predicting housing market performance Authors: Ahmed KhalafallahDate of publication: October,2008Methodology: Artificial Neural network-based model

Summary:The main objective of this paper is to present the development of ANN models that would help real estate investors and home developers predict the change in house prices in the short-term. To this end, the present model is designed to provide a number of unique and practical capabilities, including:
(1) utilizing artificial neural networks in order to build generalized knowledge about the past and current performance of housing industry and
(2) predicting the ratio between house averages sale and asking prices.

2.3 An intelligent model for predicting the sales of a productAuthors: Avinash Kumar Sharma, Neha Goel, Jatin Rajput, Mohd. BilalDate of Publication: 29-31, January 2020Methodology: Machine learning and Deep learning

Summary:The methodology used in this paper is based on machine learning and deep learning for analysing the data and fetch useful information from the patterns and trends. In this approach, they used sentimental analysis of products by setting up models in a productive way and had occupied accurate models for each case in a life cycle of a product.

IV. PROBLEM FORMULATION

The problem statements can be divided into 3 keywords, i.e., Present scenario, product update and change, and future forecast of products at a given input.

1. PRESENT SCENARIO:Present scenario will give the current situation of the product in the market.Classify the status of the product in terms of sales, price, and as an input.Existing competitors influencing the current state of the product.

2. PRODUCT CHANGE AND UPDATE:Any change in existing product or a substitute to existing product is launched in the market; it will also affect its price and sales in the market.APPROACH: The changes in inputs such as sales, price, and the present scenario will be used to forecast or predict the situation of the product.Clustering algorithm will group the data based on similarity and dissimilarity between them.Random forest Algorithm is used both for classification and regression kind of problems.

3. FUTURE FORECAST**:**Future prediction will be based on the previous data of the product. APPROACH: Previous data of sales will act as an input and future sales will be predicted.METHOD: Random Forest Tree Algorithm. It reduces over-fitting and therefore it is more accurate.

V. EXISTING SYSTEM

A sales forecast predicts what a salesperson, team, or company will sell weekly, monthly, quarterly, or annually.When a new product is launched in the market, there is no data about the product is going to perform in the market. Therefore, in order to get the data about feedback of the customers is taken in the form of the survey.These surveys help in moving the correct direction product generation.

VI. PROPOSED SYSTEM

Forecasting is needed to upgrade the products and even to tune the production of the products in the industry.In this project, our focus is on the forecasting the demands of the product using the random forest and clustering algorithms to help in growth of the business.The datasets are taken from www.kaggle.com and implemented in Python. It can be said as a self-assessment tool which uses the statistics of the past and the current sales in order to predict future performance.

VII. FEASIBILITY STUDY

The feasibility of the project is analysed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential. Three key considerations involved in the feasibility analysis are:

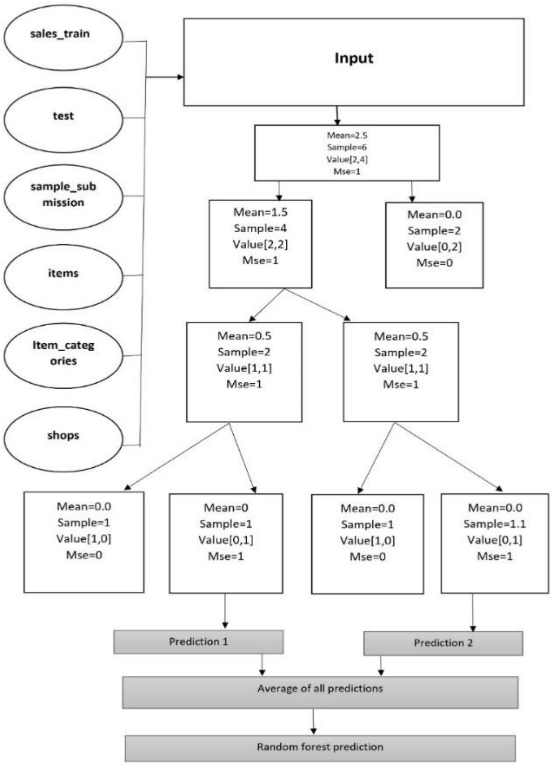
1. ECONOMIC FEASIBILITY**:**This study is carried out to check the economic impact will have on the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customised products have to be purchased.

2. TECHNICAL FEASIBILITY:This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed xix must not have a high demand on the available available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes for the implementing this system.

3. OPERATIONAL FEASIBILITY**:** The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

VIII. ARCHITECTURAL DIAGRAM

The architecture diagramcan help system designers and developers visualize the high-level, overall structure of their system or application, in order to ensure the system meets their users' needs. Using architecture diagrams, you can also describe patterns that are used throughout the design. It's somewhat like a blueprint that you use as a guide so that you and your colleagues can discuss, improve, and follow. This Architecture Diagram explains the workflow of the overall project. The datasets are taken from Kaggle platform and the datasets were put together and implemented using python libraries. By using Random Forest Algorithm method all the datas are combined and "n" number of predictions are given. By taking average of all the predictions the correct and accurate Random Forest Algorithm is generated.


sales_train
test
sample_sub mission
items
Item_categ ories
shops
Input
Mean=2.5 Sample=6 Value[2,4] Mse=1
Mean=1.5 Sample=4 Value[2,2] Mse=1
Mean=0.0 Sample=2 Value[0,2] Mse=0
Mean=0.5 Sample=2 Value[1,1] Mse=1
Mean=0.5 Sample=2 Value[1,1] Mse=1
Mean=0.0 Sample=1 Value[1,0] Mse=0
Mean=0 Sample=1 Value[0,1] Mse=1
Mean=0.0 Sample=1 Value[1,0] Mse=0
Mean=0.0 Sample=1.1 Value[0,1] Mse=1
Prediction 1
Prediction 2
Average of all predictions
Random forest prediction

The above fig 5.1 is the architectural diagram for sales prediction

SYSTEM 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.

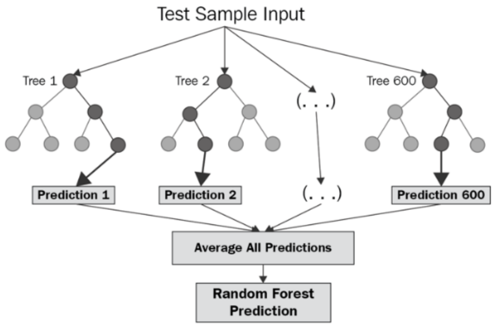
MODEL USED FOR GENERATING THE FORECAST:

Before deciding the forecasting model, dataset is required. Clustering method is used to cluster the items showing similar patterns in their behaviour.

Further, data analysis is done using Random Forest Algorithm and Neural Network.

RANDOM FOREST ALGORITHM:

Random forest comprises of many decision trees. It is a supervised algorithm used for regression to predict the future demand.



Representation of Random Forest Algorithm

The Random forest algorithm, first, processes the training set to develop a random forest model, and then produce K decision trees to set up a “forest”. Additionally, the algorithm categorizes the decision tree and generates a vote. The decision tree with maximum number of votes is the prediction.

ADVANTAGES:

1. Random forest works great with multi-dimensional data.
2. Random forest can deal with outliers.
3. This algorithm has a high learning rate as compared to others.
4. Random forests is robust.
5. It can be useful for both classification and regression type of problems.

DATASET:

We are provided with daily historical sales data. The task is to forecast the total amount of products sold in every shop for the test set. Note that the list of shops and products slightly changes every month. Creating a robust model that can handle such situations is part of the challenge.

FILE DESCRIPTIONS:

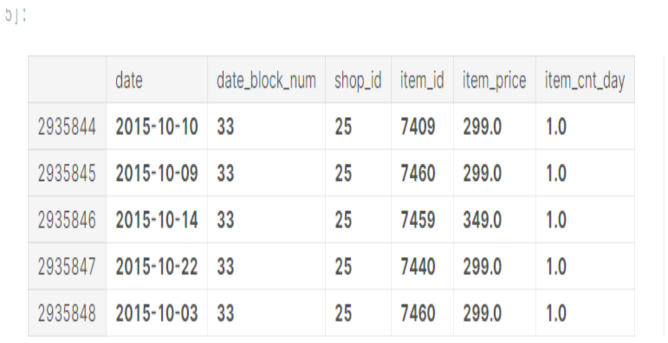
- sales_train.csv - the training set. Daily historical data from January 2013 to October 2015.
- test.csv - the test set. You need to forecast the sales for these shops and products for November 2015.
- sample_submission.csv - a sample submission file in the correct format.
- items.csv - supplemental information about the items/products.
- item_categories.csv - supplemental information about the item’s categories.
- shops.csv- supplemental information about the shops.

DATA FIELDS:

- ID - an Id that represents a (Shop, Item) tuple within the test set
- shop_id - unique identifier of a shop
- item_id - unique identifier of a product
- item_category_id - unique identifier of item category
- item_cnt_day - number of products sold. You are predicting a monthly amount of this measure
- item_price - current price of an item
- date - date in format dd/mm/yyyy
- date_block_num - a consecutive month number, used for convenience. January 2013 is 0, February 2013 is 1, ..., October 2015 is 33
- item_name - name of item
- shop_name - name of shop
- item_category_name - name of item category

DATA STRUCTURE EXPLORATION:

The csv files are loaded to pandas data frame. There are 2935849 training examples and 214200 test samples. There are 2935849 days of record for 33 months in train data frame Last 5 records are shown below



	date	date_block_num	shop_id	item_id	item_price	item_cnt_day
2935844	2015-10-10	33	25	7409	299.0	1.0
2935845	2015-10-09	33	25	7460	299.0	1.0
2935846	2015-10-14	33	25	7459	349.0	1.0
2935847	2015-10-22	33	25	7440	299.0	1.0
2935848	2015-10-03	33	25	7460	299.0	1.0

CATEGORY SPLIT AND TRANSLATION:

The data originates from Russian stores so it also contains names, categories written in cyrylic. For further processing we split the item_category_name column into category and subcategory.

DATA CLEANING AND PREPARATION:

We will join the data into one data frame

```
train = pd.merge(train, shops, on='shop_id', how='left')
train = pd.merge(train, items, on='item_id', how='left')
train = pd.merge(train, item_cats, on='item_category_id', how='left')
```

We drop the duplicate samples from training set to avoid overfitting the model with same samples.

Since the problem states to predict the sales for the samples in testing set, we need to train models with samples among the unique shop id and item id. Hence, we keep the training samples which are only testing set.

```
print('train:',train.shape)
train.drop_duplicates(subset=['date','date_block_num','shop_id','item_id','item_cnt_day'],inplace=True)
shops_test=test.shop_id.unique()
items_test=test.item_id.unique()
train=train[train.shop_id.isin(shops_test)&train.item_id.isin(items_test)]
print('train:',train.shape)
```

Check if we have some missing values in columns.

```
train.isnull().sum()
```

The most expensive item in dataset belongs to category Программы - Длядома и офиса -> Programs - Home and Office Item name: Radmin 3 - 522 лиц. ->Radmin 3 - 522 persons. Radmin 3 is a program providing remote access to computers in a network. It looks like itsone shot sale and order was for 522 individual licenses. We will confirm this row as outlier with IQR method.

```
fig,ax=plt.subplots()
ax.set(yscale="log")
sns.boxplot(y=train.item_price,whis=1.5,ax=ax)
```

Clearly the item_price value grater than 100000 is an outlier

```
train=train[~(train.item_price>100000)]
```

Now we will take a closer look at number of products sold. Looking at output of describe we noticed that percentiles and max value of sales could indicate that we have an outlier above value of 1000 sales per day. Additionally min value is negative so we will also take a look at these rows

```
train[(train.item_cnt_day>1000)]
```

Unfortunately we don't know what negative sales data mean. It could be just an error. Taking a look at product types we also suspect that stolen items could be indicated by negative values. We will take an assumption that it is a theft and we want to predict only sales. This brings us to conclusion that we should drop these rows

```
train=train[~(train.item_cnt_day<0)]
```

RANDOM FOREST CLASSIFIER:

The random forest classifier is one of the ensemble algorithm and is known for it's robust performance.

```
random_forest=RandomForestClassifier(n_estimators=100)
random_forest.fit(train_cleaned_df.iloc[:,(train_cleaned_df.columns!=33)].values,train_cleaned_df.iloc[:,train_cleaned_df.columns==33].values
```

- **Random Forest Score:**

The mean square error on the training set is 0.00733900068647

```
preds=random_forest.predict(train_cleaned_df.iloc[:,(train_cleaned_df.columns!=33)].values)
rmse=np.sqrt(mean_squared_error(preds,train_cleaned_df.iloc[:,train_cleaned_df.columns==33].values))
print(rmse)
```

IX. EXPERIMENTAL RESULTS

Count for different items categories:

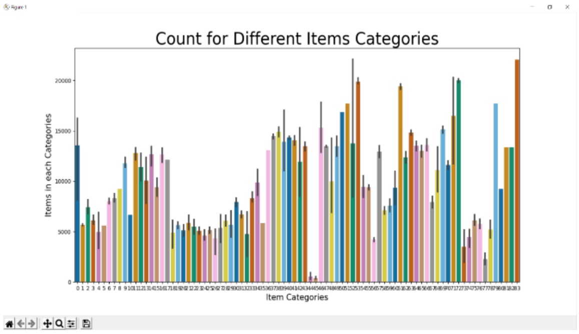


Fig: Count for different items categories

Plot above shows item categories for each month. First obvious observation is that items in each categories increases at the end of the year. Second observation is that overall item categories.

Date blocks according to months:

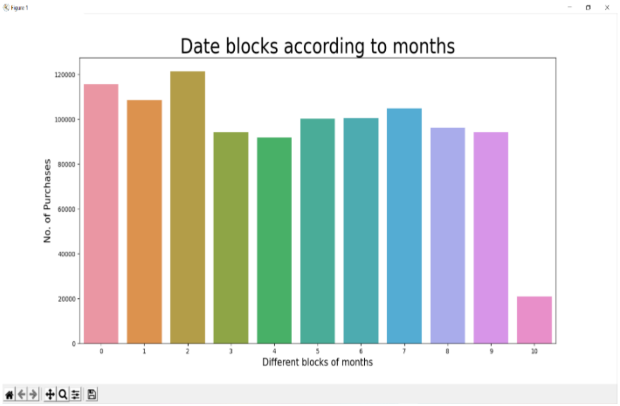


Fig: Date blocks according to months

Above plot shows that the Number of Purchases in each month. The first observation is the Number of Purchases and the second observation is the different blocks of months

The busiest months for the shops:

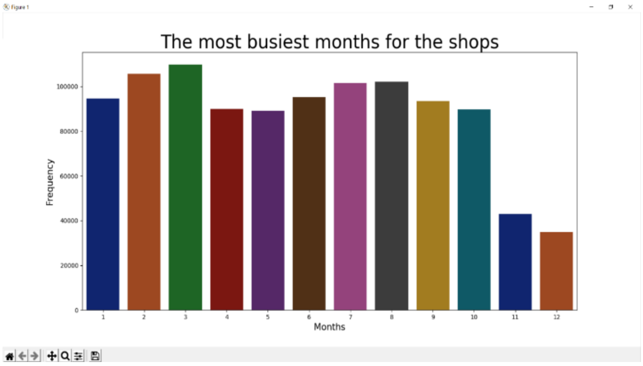
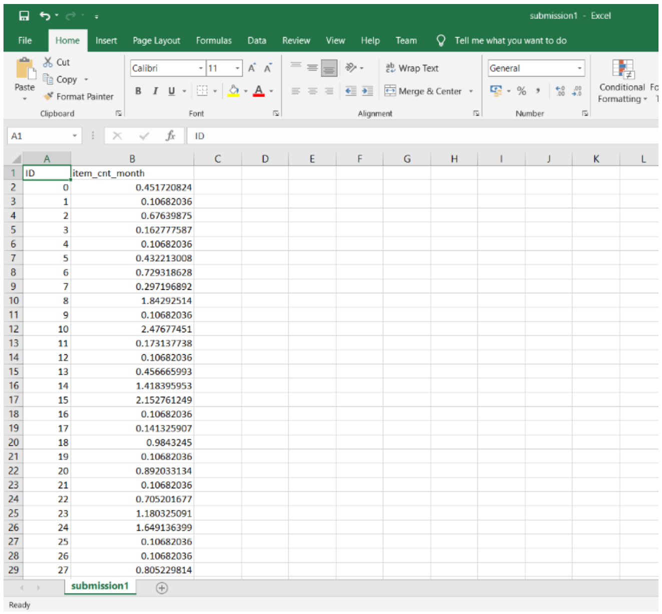


Fig: The busiest months for the shops

The above-mentioned figure is the graphical representation of frequency of sales in particular shops in particular months

PREDICTION OF SALES:

submission1.psv:



	A	B
1	ID	item_cnt_month
2	0	0.451720824
3	1	0.10682036
4	2	0.67639875
5	3	0.162777587
6	4	0.10682036
7	5	0.432213008
8	6	0.729318628
9	7	0.297196892
10	8	1.84292514
11	9	0.10682036
12	10	2.47677451
13	11	0.173137738
14	12	0.10682036
15	13	0.456665993
16	14	1.418395953
17	15	2.152761249
18	16	0.10682036
19	17	0.141325907
20	18	0.9843245
21	19	0.10682036
22	20	0.892033134
23	21	0.10682036
24	22	0.705201677
25	23	1.180325091
26	24	1.649136399
27	25	0.10682036
28	26	0.10682036
29	27	0.805229814

The submission1.psv folder consist of the prediction of future sales of various items in various shops.

X. CONCLUSION AND FUTURE ENHANCEMENT

Sales forecasting has become highly sophisticated and plays a vital role to the competitiveness of many companies.We have tried to deploy different learning algorithms such as Random forest algorithm used to implement products' demand forecasting to improve an organization's profit has performed best in achieving our objective in this dataset.This can be attributed to the good feature engineering and the ability to try out a large range of hyperparameters during optimization. Future work for this is involves comparing the performance evaluation results of the chosen regression techniques obtained from this thesis with the results obtained from deep learning methods which could help the researchers in getting better trends and results. As mentioned earlier, due to the lack of information about external or environmental variables like promotions, discounts, etc. Such variables are not considered in the modeling and sales forecasting experiment. It is suggested to use the related data of the company for further research to obtain more accuracy.

REFERENCES

[1] Predicting Future Sales, Kaggle URL=https://www.kaggle.com/c/competitive-data-science-predict-future-sales

[2] "Random Forest Algorithm", URL= https://towardsdatascience.com/the-mathematical-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df7e3

[3] Mining the past to determine the future market: Sales forecasting using TSDM framework, AngliqueD.Lacasandile, Jamin d. Niguidula, Jonathan M. Caballero, November 5-8,2017

[4] Neural network-based model for predicting housing market performance, Ahmed Khalafallah, October,2008

[5] An intelligent model for predicting the sales of a productAvinash Kumar Sharma, Neha Goel, Jatin Rajput, Mohd. Bilal, 29-31, January 2020