AUTOMATION OF BEST-FIT MODEL SELECTION USING A BAG OF MACHINE LEARNING LIBRARIES FOR SALES FORECASTING

Pauline Sherly Jeba P[1], Manju Kiran[1],
Amit Kumar Sharma[1], Divakar Venkatesh[2]

[1]Engineering Data Science, RBEI, Robert Bosch GmbH, India
[2]Enterprise Solutions-SAP, Robert Bosch GmbH, Postfach, Stuttgart, Germany

ABSTRACT

Sales forecasting became crucial for industries in past decades with rapid globalization, widespread adoption of information technology towards e-business, understanding market fluctuations, meeting business plans, and avoiding loss of sales. This research precisely predicts the automotive industry sales using a bag of multiple machine learning and time series algorithms coupled with historical sales and auxiliary features. Three-year historical sales data (from 2017 till 2020) were used for the model building or training, and one-year (2020-2021) predictions were computed for 900 unique SKU's (stock-keeping units). In the present study, the SKU is a combination of sales office, core business field, and material customer group. Various data cleaning and exploratory data analysis algorithms were implemented over raw datasets before use for modeling. Mean absolute percentage error (mape) were estimated for individual predictions from time series and machine learning models. The best model was selected for unique SKU's as per the most negligible mape value.

KEYWORDS

Automotive, Machine learning, Time series, Artificial Intelligence, Sales-forecasting, Cognitive approach, Outliers.

1. INTRODUCTION

Sales forecasting is to ease the forecasting process and to create a convenient and easy-to-use application/tool for a business [1]–[3]. A sales forecast indicates how much of a particular product is likely to be sold in a specified future period in a specified market at a specified price [4]. Sales forecasting arranges in advance for raw materials, equipment’s, labour etc. Accurate sales forecasting is one of the most essential requirements of any organization supply chain [5], [6]. It requires developing a unique model that pre-processes the input data and ensembles the output of two parallel advanced forecasting engines that use state-of-the-art machine learning, deep learning algorithms, and time series algorithms to generate future sales forecasts [7], [8].

Sales forecasting adds value across an organization. Finance, for example, relies on forecasts to develop budgets for capacity plans and hiring [9]. The production uses sales forecasts to plan raw materials [10]. Forecasts help sales operations with territory and quota planning, supply chain with material purchases [6], [11]. Sales forecast helps to improve the decision-making about the future in terms of reduction of sales pipeline and forecast risks [12].

DOI: 10.5121/ijaia.2021.12602

Sales forecasting can be sub-categorized into three categories: (i) short term forecasting (for a period up to 3 months ahead), (ii) medium-term forecasting (for one year ahead, its crucial for business budgeting), and (iii) long-term forecasting (for three years, its crucial for long term resource implications) [13].

Decision-makers rely on these forecasts to plan for business expansion and to determine how to fuel the company growth [14]. Therefore, in many ways, sales forecasting affects every department in the organization:

- A sales forecast helps every business to make a better business decision. It helps overall in business planning, budgeting, and risk management.
- Sales forecasting allows companies or industries to efficiently allocate resources for the future growth and manage their cash flow.
- Sales forecasts help sales teams to achieve their goals by identifying an early warning signals in their sales pipeline and to course-correct before it’s too late
- Sales forecasting helps businesses estimate their costs and revenue accurately based on which they can predict their short-term and long-term performance.

Having an accurate sales forecast is still a significant challenge in front of most of the companies. Traditional forecasting methods based on various intuition; companies end up having poor visibility into projected sales. When a company continuously misses its sales forecast, it can have a negative impact on its valuation over the long-time duration. Exceeding your forecasts isn’t good news either. When companies cannot estimate how much revenue it will generate accurately, they cannot hire or invest in keeping with the growth, which could lead to several missed opportunities [15].

Our technique uses data-driven statistical methods to clean the data of any potential errors or outliers and impute missing values if any. Once the forecast is generated, it is post-processed with Seasonality and Trend corrections, if required. Since the final forecast results from a statistically pre-validated ensemble of multiple models, the forecasts are stable, and accuracy variation is minimal across periods and forecast horizons. Hence, it is better at estimating future sales than conventional techniques.

2. DATASET AND METHODOLOGY

This section elaborated the dataset used for the study and implemented methodology towards precise sales forecasting.

2.1. Dataset

Input dataset used for the sales forecasting analysis were subcategorized into two sections:

(i) Baseline data and
(ii) External data.

Figure 1 shows the input data categorization for the sales forecasting perspective. The baseline data represents the actual sales quantity for various SKU combinations. External data represents additional information which may impact direct or indirect sales. We used price change data, secondary sales (retailer sales data), schemes, and policy as external datasets.

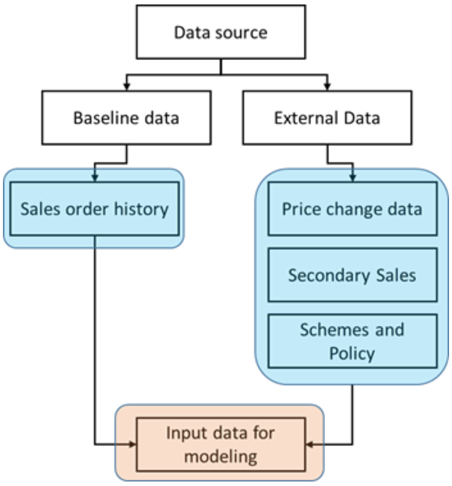

Figure 1. Input data categorization for the modeling.

2.2. Problem Definition

(i) Problem Definition: Identifying the main business goals and to set expectations before any development phase.
(ii) Collecting Information: Integrate the different data sources that will be used as the foundation for the models, and data needs to be effectively interpreted and analysed.
(iii) Exploratory data analysis: Primary analyses is carried out on data to use insights from results to define further steps and assessment to be done to find the relation between features in the dataset.
(iv) Machine Learning: The process of applying statistical algorithms on the prepared dataset, providing a rigorous framework to test those models, and insights drawn from the previous phase are used to choose the most appropriate models that could be applied.
(v) Validation & Testing: Assess models' accuracy and robustness, and models should be generalized and be able to produce reliable results outside of the dataset they have been developed.
(vi) Result Communication: Communicate the advanced analytics models results effectively and translate them into actionable business insights. Model results should assist businesses in decision-making.

2.3. Methodology

Figure 2 represents the novel methodology used in sales forecasting. The Sales forecasting methodology was subdivided into multiple steps like:

(i) Customized classification of parts according to movement, revenue, business criticality, and sale volume,
(ii) Visualization of sales forecast across the inventory along with KPIs (key per-forming indicator) such as over prediction and under prediction estimates and revenue/business impact,
(iii) Visibility to your inventory holding costs and insufficient inventory costs (backorder / loss of sale) and month on month comparison for better insights and decisions,
(iv) Drill down from region, plant, market till SKU level, comparison across SKU's for identifying patterns of sale, the impact of one SKU's over other,
(v) Liberty to the user to select preferred algorithm to view the forecast, option to update the hyperparameters and manually override in the forecast value,

(vi) Configuration of custom rules for SKU's according to business/domain,

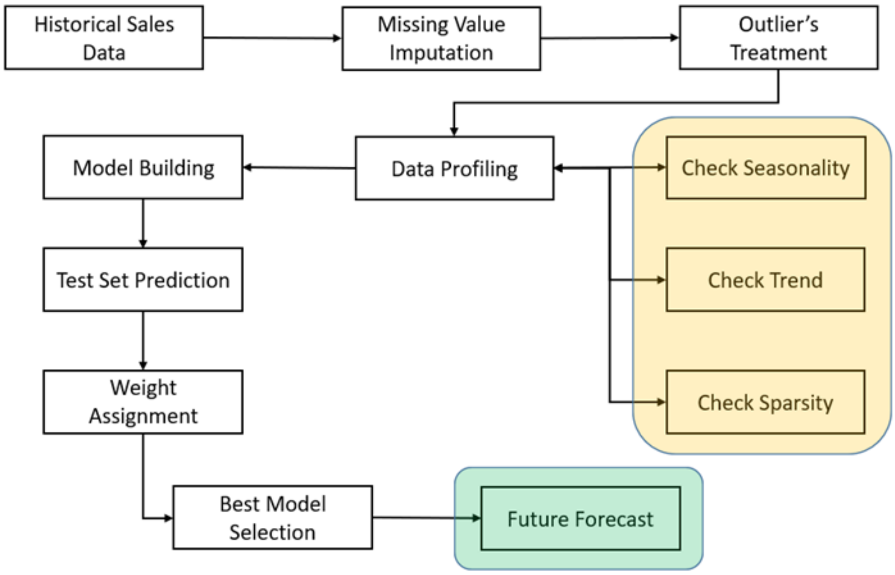

Figure 2. Methodology for the sales forecasting

(i) SKU details along with AI (Artificial Intelligence) driven material insights for a better understanding of the sku's
(ii) An established pipeline for considering external factors in forecasting de-mand,
(iii) Robust models choosing the hyperparameters accordingly and tuning it to forecast and take care of the stability of the models,
(iv) Advanced methods to deal with the bulk order, declining products, sister SKU mapping (predecessor and successor linkages), and seasonal products.

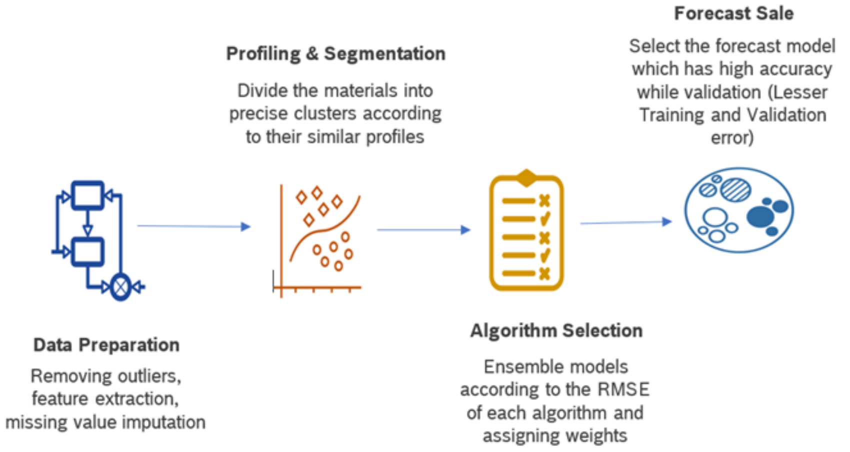
Forecast Sale
Select the forecast model which has high accuracy while validation (Lesser Training and Validation error)
Profiling & Segmentation
Divide the materials into precise clusters according to their similar profiles
Data Preparation
Removing outliers, feature extraction, missing value imputation
Algorithm Selection
Ensemble models according to the RMSE of each algorithm and assigning weights

Figure 3. Methodology process overview

Figure 3 represents the methodological overview as four major buckets to-wards precise sales forecasting. The data preparation, profiling, segmentation, algorithm selection, and forecast sale are the main four buckets or sub-sections of complete methodology.

3. RESULTS AND DISCUSSION

3.1. Data pre-processing

For our analysis, we extracted SKUs from the Primary and Secondary Sales History data. The SKUs are drilled down to their Sales Office, core business field, and the material group level. The Net value for each unique SKU from 2017 January till 2020 March is taken as the Training Data. It also includes external factors which would cause an impact on the sales forecast, thus making it more reliable and accurate.

3.2. Data Merge

The external and the base data required for training were available in different business files. These files were merged using merge functions available in pandas. The data in each file were combined together based on the specified granularity level (Sales office, core Business field, and Material customer group) along with the provided Billing date for each product.

3.3. Data Transformation

The sales history data can be any interval type - daily, weekly, monthly, yearly, etc. Aggregate the training data based on the required Forecast interval type. In our approach, we are forecasting sales on a monthly level.

3.4. Data Imputation

3.4.1. Missing Value Imputation

Suppose the count of the missing values is greater than a certain threshold. In that case, missing values are imputed with zero. Otherwise, the neighbouring data points and the data points from corresponding time frames in the previous or future periods, if available, are utilized to impute the missing values. The covid affected months are imputed with the average of the Net value from the previous quarter.

3.4.2. Missing Dates Imputation

Based on the identified interval type in the training data, Missing periods/dates are imputed. The missing value of the imputed periods are taken care by the above method.

3.5. Data Pre-Processing/Data Cleaning

3.5.1. Exploratory Data Analysis (EDA)

Univariate Analysis: A statistical analysis of the input data is performed. Mean, Median, Standard deviation, Minimum, and Maximum is checked.

SKUs with Negative values in the target column can be discarded as they are treated as an outlier.

Pearson Correlation: Pearson’s correlation is a statistical method to measure the strength of relationship between two variables and their association.

According to the correlation result, Features that are highly correlated to the target column (‘Net value’) are chosen.

All the features with Pearson's correlation>=0.5 are chosen while the others are dropped.

Collinearity Analysis: We check for Multicollinearity in the features to identify if more than two explanatory variables/features are more linearly related to each other.

We check for the Variance Inflation Factor (VIF), the features with Variance Inflation Factor > 10 indicate multicollinearity. Such highly correlated features can be dropped as it can lead to skewed or misleading results.

3.5.2. Outliers Treatment

Outlier Treatment is applied for a bucket of data. Bucket size is selected based on the interval type. For example, outlier treatment is applied for a span of 12 months if it is monthly data. Median Absolute Deviation or Mean Standard Deviation is performed on the data. Data points which are more than three times of standard deviations from the mean are clipped off.

3.5.3. Determination of Lags

ACF or Auto Correlation Plots are used to determine the lags. It describes how well the present value of the series is related with its past values. The Partial Auto Correlation Plots (PACF) gives a stationary time series partial correlation with its own lagged values. It describes the direct relationship between an observation and its lag. Optimal values of parameters “p”, “d”, and “q” are thus calculated from the ACF (Auto Correlation Function) and PACF plots. If both ACF and PACF decrease gradually, the time series has to be made stationary using “d”, the degree of difference.

3.5.4. Check Stationary

Three basic criterion criteria classify data as stationary:

The mean of the series should not be a function of time. It should be constant.

The variance of the series should not be a function of time. This is called homoscedasticity.

The covariance of the i^{th} term and the $(i+m)^{th}$ term should not be a function of time.

The data is made stationary by differencing the dataset using lag. The stationarity of dataset is verified using Dickey-Fuller Test.

3.6. Best Fit Model Selection

Data Profiling

Before identifying the best-fit model, the prepared dataset undergoes a method called profiling. Data profiling is a method of examining or understanding the characteristics of the dataset. As represented in figure 4, data profiling, the characteristics such as Seasonality, Trend, Sparsity are detected for each SKU with which they are grouped under each model cluster. The models are clustered under different groups to avoid high computational time. Training data is further split into Train and validation set on a ratio of 7:3.

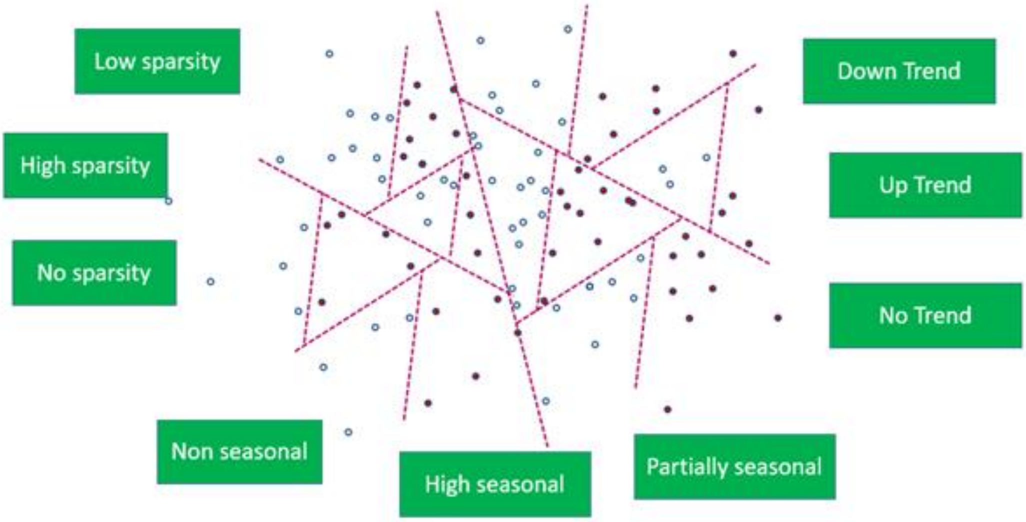

Figure 4. Profiling based on Sales Characteristics.

3.7. Model Building

3.7.1. Time Series Algorithms

ARIMA (Auto Regression Integrated Moving Average), ARMA (Auto Regression Moving Average), AR (Auto Regression), Moving Average, Weighted Moving Average (WMA), Exponential Time Smoothing, Holt-Winters, and various other models are run on the datasets. A total of about eight-time series models were used. If the data is highly sparse or insufficient, weighted moving average is performed. RMSE (Root Mean Square Error) is calculated for the validation set run on each model. The top 3 models with minimum RMSE value are chosen as the best Time series models.

3.7.2. Machine Learning Algorithms

Various machine learning algorithms like Linear Regression, Decision Tree Regression and SVM (Support Vector Machine) are run on the datasets. A total of about 13 different machine learning models are used. The Hyperparameters are tuned using a Grid search and random search approach. The top 3 models with minimum RMSE value are chosen as the best Machine learning models.

3.7.3. Ensemble Algorithms

The errors calculated from the best models of Time Series and Machine Learning algorithms are used to calculate the weights. The below mentioned formula calculates the weights:

$$w_{ts} = {}^{1/error_t}\Big/\left({}^{1}/_{error_t} + {}^{1}/_{error_m}\right)$$

$$w_{ml} = {}^{1/error_m}\Big/\left({}^{1}/_{error_t} + {}^{1}/_{error_m}\right)$$

These calculated weights are multiplied with respective machine learning and time series forecast results and by ensembling, the final forecasts are calculated.

3.8. Performance Metrics & Results

Finally, the forecast values and error RMSE of the validation set for the best TS, ML and ensemble algorithms are considered. The approach which has the minimum RMSE is opted to Forecast the data for the future period of time. The Forecast accuracies of the TS, ML and Ensemble method are compared. The performance metrics suited well to compare the 3 different approaches are Forecast Accuracy and MAPE. The forecast accuracies for the five consecutive months are compared to check their stability.

Table 1 represents the accuracy (%) for the various ML and TS algorithms for a SKU. Table 2 shows the weights for the TS and ML validation sets.

Table 1. Test set accuracies (%) of each model along the Forecast accuracy of new data

SKU	ADA (test set)	RF (test set)	XGB (test set)	ARIMA (test set)	NAÏVE (test set)	HWES (test set)	ENSEMBLE (test set)	FACC (New Forecast)
SKU_ID	87.0	74.2	38.3	39.6	5.6	43	60.5	95.8

Table 2. Calculated weights for the ML and TS validation sets

Calculated Weights ML (Validation Set)	Calculated Weights TS (Validation Set)
0.3970	0.6025

The top best Models identified in ML were ADA-BOOST, RANDOM FOREST and XGBOOST and the top best Models identified in TS were ARIMA, NAÏVE, HWES (Holt-Winters exponential Smoothing). An Ensemble approach is applied on the first best model of ML and TS by calculating their Weights. The model/approach with the least minimum error in the test set is used to forecast the future period of time. Figure 5 represents the training and prediction outcomes for a sample SKU.

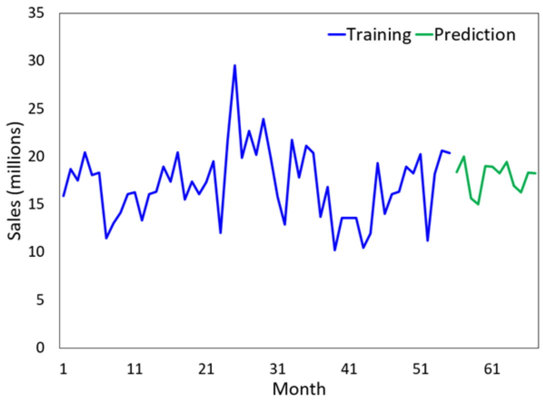

Figure 5. Training and prediction representation for a sku. Here we used 55 months training dataset to predict next 12 months.

4. CONCLUSION

During the process of sales forecasting through machine learning and time series algorithm implementation, the sku's with more than 12 months of historical sales data machine learning algorithm tend to consistently provide higher forecasting accuracy compared to the time series model. With dataset less than 12 months, it was decided to perform time series forecasting model. Sales forecasting precision can further be improved by exploring deep learning methods, which can be implemented in the future. External market data like GDP (Gross Domestic Product), growth, fuel price, weather (parts like viper), etc. can be consider in future.

Certain automotive aftermarket components with sparse data ML and TS models were not suitable, hence best method for the sales forecasting was found to be traditional simple moving average of historical sales or naïve approach.

ACKNOWLEDGEMENTS

The authors would like to thank our organization leadership team for supporting and guiding us throughout the project.

REFERENCES

[1] B. Seaman and J. Bowman, "Applicability of the M5 to Forecasting at Walmart," *Int. J. Forecast.*, 2021.

[2] R. J. Kuo and K. C. Xue, "A decision support system for sales forecasting through fuzzy neural networks with asymmetric fuzzy weights," *Decis. Support Syst.*, vol. 24, no. 2, pp. 105–126, 1998.

[3] Z. P. Fan, Y. J. Che, and Z. Y. Chen, "Product sales forecasting using online reviews and historical sales data: A method combining the Bass model and sentiment analysis," *J. Bus. Res.*, vol. 74, pp. 90–100, May 2017, doi: 10.1016/J.JBUSRES.2017.01.010.

[4] G. Kulkarni, P. K. Kannan, and W. Moe, "Using online search data to forecast new product sales," *Decis. Support Syst.*, vol. 52, no. 3, pp. 604–611, Feb. 2012.

[5] T. Boone, R. Ganeshan, A. Jain, and N. R. Sanders, "Forecasting sales in the supply chain: Consumer analytics in the big data era," *Int. J. Forecast.*, vol. 35, no. 1, pp. 170–180, Jan. 2019.

[6] V. Sohrabpour, P. Oghazi, R. Toorajipour, and A. Nazarpour, "Export sales forecasting using artificial intelligence," *Technol. Forecast. Soc. Change*, vol. 163, p. 120480, Feb. 2021, doi: 10.1016/J.TECHFORE.2020.120480.

[7] D. Mezzogori and F. Zammori, "An entity embeddings deep learning approach for demand forecast of highly differentiated products," *Procedia Manuf.*, vol. 39, pp. 1793–1800, 2019.

[8] E. Spiliotis, S. Makridakis, A. Kaltsounis, and V. Assimakopoulos, "Product sales probabilistic forecasting: An empirical evaluation using the M5 competition data," *Int. J. Prod. Econ.*, vol. 240, p. 108237, Oct. 2021, doi: 10.1016/J.IJPE.2021.108237.

[9] P. Ramos, N. Santos, and R. Rebelo, "Performance of state space and ARIMA models for consumer retail sales forecasting," *Robot. Comput. Integr. Manuf.*, vol. 34, pp. 151–163, 2015.

[10] P. C. Chang, Y. W. Wang, and C. Y. Tsai, "Evolving neural network for printed circuit board sales forecasting," *Expert Syst. Appl.*, vol. 29, no. 1, pp. 83–92, Jul. 2005.

[11] T. Boone, R. Ganeshan, A. Jain, and N. R. Sanders, "Forecasting sales in the supply chain: Consumer analytics in the big data era," *Int. J. Forecast.*, vol. 35, no. 1, pp. 170–180, Jan. 2019, doi: 10.1016/J.IJFORECAST.2018.09.003.

[12] J. Mun, "A Primer on Quantitative Risk Analysis," *Multi-Asset Risk Model.*, pp. 63–118, 2014, doi: 10.1016/B978-0-12-401690-3.00003-2.

[13] S. Smyl, "A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting," *Int. J. Forecast.*, vol. 36, no. 1, pp. 75–85, Jan. 2020.

[14] Y. Chen, H. Zhao, and L. Yu, "Demand forecasting in automotive aftermarket based on ARMA model," *2010 Int. Conf. Manag. Serv. Sci. MASS 2010*, pp. 10–13, 2010, doi: 10.1109/ICMSS.2010.5577867.

[15] "Erratum regarding missing Declaration of Competing Interest statements in previously published

articles (International Journal of Forecasting (2019) 35(1) (170–180), (S0169207018301523), (10.1016/j.ijforecast.2018.09.003)),” *Int. J. Forecast.*, vol. 37, no. 3, pp. 1310–1311, Jul. 2021.

AUTHORS

Pauline Sherly Jeba P is a Data Scientist at RBEI Robert Bosch GmbH, She is having more than 4 years of experience in demand forecasting, energy analytics and connect products data domains.

Manju Kiran B A is a Program Manager at RBEI Robert Bosch GmbH, Manju is having more than 12 years of experience in data science domain.

Dr. Amit Kumar Sharma is a Data Scientist at RBEI Robert Bosch GmbH, Amit is having more than 10 years of experience in multinational and multi-dimensional projects.

Divakar Venkatesh is a Senior Architect at SAP, Robert Bosch GmbH, Divakar is having 17 years of experiencing in implementing SAP to CRM, sales and demand forecasting solutions for global customers.