IJSETT International Journal

"International Journal for Science and Emerging Technologies with Latest Trends"
4(1): 25-24 (2012)
ISSN No. (Online) : 2250-3641, ISSN No. (Print):2277-8136

Predicting the future of car manufacturing industry using Naïve Bayse Classifier

Sukhmeet Kaur
Assistant Professor,
Deptt of IT
CEM,Kapurthala
Punjab ,India

Kiran Jyoti
Assistant Professor,
Deptt of IT
GNDEC,Ludhiana
Punjab ,India

Abstract

Data mining is a process that discovers interesting information from the hidden data which can either be used for future prediction and/or intelligently summarizing the details of the data. The applications of data mining in the field of predicting the future of any car manufacturing industry is gaining a lot of research interest now days. There are a number of research workers interested in this field. Among various data mining techniques regression, decision trees and naïve bayse algorithm is used for the prediction purposes. In this paper, naïve bayse algorithm is used for the prediction of future of number of cars which is useful for the car manufacturing industry. The prediction results are compared with the actual and real world values in order to validate the results obtained using naïve bayes alogorithm.

KEYWORDS

Data mining, Predictive Modeling, naïve bayse.

I. Introduction

The automotive industry in India is one of the largest industries in the world and one of the fastest growing globally. India's passenger car and commercial vehicle manufacturing industry is the sixth largest in the world, with an annual production of more than 3.9 million units in 2011.[1] According to recent reports, India overtook Brazil and became the sixth largest passenger vehicle producer in the world (beating such old and new auto makers as Belgium, United Kingdom, Italy, Canada, Mexico, Russia, Spain, France, Brazil), growing 16 to 18 per cent to sell around three million units in the course of 2011-12.[2] The majority of India's car manufacturing industry is based around three clusters in the south, west and north. The southern cluster consisting of states like

Chennai and Bangalore is the biggest with 35% of the revenue share. The western hub near Mumbai and Pune contributes to 33% of the market and the northern cluster round the National Capital Region contributes32%.

The Indian Automobile Industry manufactures over 11 million vehicles and exports about 1.5 million each year.[18] The dominant products of the industry are two-wheelers with a market share of over 75% and passenger cars with a market share of about 16%.Today, India is among the world's largest producers of small cars. The New York Times has rated India as a very strong engineering base with an incomparable expertise in the arena of manufacturing a number of low-cost, fuel-efficient cars has encouraged the expansion plans of the manufacturing facilities of a number of automobile leaders like Hyundai Motors, Nissan, Toyota, Volkswagen and Suzuki.India is soon becoming a hub for car manufacturers not only to sell their small cars but

IJSETT International Journal
"International Journal for Science and Emerging Technologies with Latest Trends"
4(1): 25-24 (2012)
ISSN No. (Online) : 2250-3641, ISSN No. (Print):2277-8136

also to sell the large cars. India comes a close second to China when it comes to the fastest growth in the Automobile sector in the world. There are some challenges which the future car industry's are going to face .That are less loyal customers , the need to improve productivity, the demand for producing low cost mass market vehicles and the maintenance of small top end vehicle market. This paper aims to predict the future of car manufacturing company by using data mining techniques. Especially we aim at finding the number of cars to be manufactured by a car manufacturing company by using the previous year's data. A formula based on the current data available, historical trends, and projections is used to estimate the total number of cars to be produced in a particular year. This formula is used to predict the future of car manufacturing industries in order to foretell what is going to happen in the future.

II. Data Mining

Data mining is the process of extracting knowledge hidden from large databases. The knowledge must be new and one must be able to use it. Knowledge discovery differs from traditional information retrieval from databases. In traditional DBMS, database records are returned in response to a query; while in knowledge discovery, what is retrieved is not explicit in the database. Rather, it is implicit patterns. The Process of discovering such patterns is termed data mining. Data mining finds these patterns and relationships using data analysis tools and techniques to build models. There are two main kinds of models in data mining. One is predictive models, which use data with known results to develop a model that can be used to explicitly predict values. Another is descriptive models, which describe patterns in existing data. It is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. . Data mining has not only applied effectively in business environment but also in other fields such as weather forecast, medicine, transportation, healthcare, insurance, government and etc. Data mining brings a lot of advantages when using in a specific industry. The main purpose of data mining is to extract patterns from the data at hand, increase its intrinsic value and transfer the data to knowledge.

A. Predictive Modelling

Data mining tasks can be categorized into either prediction or description. Descriptive mining techniques are Clustering, Association Rule Mining (ARM) and Sequential pattern mining. The predictive mining techniques are Classification, Regression and Deviation detection. Predictive modeling is a process used in predictive analytics to create a statistical model of future behavior. Predictive analytics is the area of data mining concerned with forecasting probabilities and trends. A predictive model is made up of a number of *predictors*, which are variable factors that are likely to influence future behavior or results. In marketing, for example, a customer's gender, age, and purchase history might predict the likelihood of a future sale.

Predictive analytics encompasses a variety of statistical techniques from modeling, machine learning, data mining and game theory that analyze current and historical facts to make predictions about future events.[1][2].In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of

IJSETT International Journal
"International Journal for Science and Emerging Technologies with Latest Trends"
4(1): 25-24 (2012)
ISSN No. (Online) : 2250-3641, ISSN No. (Print):2277-8136

conditions, guiding decision making for candidate transactions. In predictive modeling, data is collected for the relevant predictors, a statistical model is formulated, predictions are made and the model is validated (or revised) as additional data becomes available. The model may employ a simple linear equation or a complex neural network, mapped out by sophisticated software. Generally, the term predictive analytics is used to mean predictive modeling, "scoring" data with predictive models, and forecasting. However, people are increasingly using the term to describe related analytical disciplines, such as descriptive modeling and decision modeling or optimization. These disciplines also involve rigorous data analysis, and are widely used in business for segmentation and decision making, but have different purposes and the statistical techniques underlying them vary.

III. Proposed Methodology

Selecting a data mining algorithm is not an easy task, it depends upon: the data we have gathered, the problem we are trying to solve, and the computing tools that are available. Naive bayse algorithm can handle multiple predictor variables for predicting the future no. of cars which is helpful to the car manufacturing industry. With the prediction of future of no. of cars month wise and company wise , car manufacturing industry can understand the market trends, the moods, and the changing consumer tastes and preferences .The methodology is shown in the following figure 1:

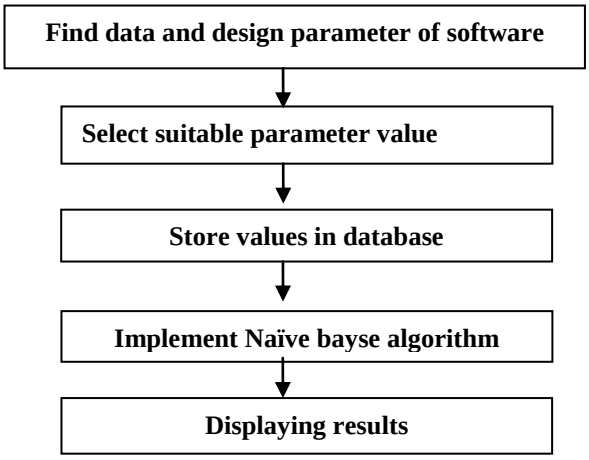

Figure 1: Flowchart of Methodology

A. Naïve bayse classifier

One highly practical Bayesian learning method is the naive Bayes learner, often called the naive Bayes classifer. In some domains its performance has been shown to be comparable to that of neural network and decision tree learning. This section introduces the naive Bayes classifier. The naive Bayes classifier applies to learning tasks where each instance x is described by a conjunction of attribute values and where the target function f (x) can take on any value from some finite set V. A set of training examples of the target function is provided, and a new instance is presented, described by the tuple of attribute values $(a_1, a_2 \ldots . a_n)$The learner is asked to predict the target value, or classification, for this new instance.

IJSETT International Journal
"International Journal for Science and Emerging Technologies with Latest Trends"
4(1): 25-24 (2012)
ISSN No. (Online) : 2250-3641, ISSN No. (Print):2277-8136

The Bayesian approach to classifying the new instance is to assign the most probable target value, v_{MAP},given the attribute values $(a_1, a_2 \ldots \ldots a_n$ that describe the instance.

$$v_{MAP} = argmax\ p\ (v_j|a_1, a_2 \ldots\ldots a_n) \qquad (1)$$

We can use Bayes theorem to rewrite this expression as

$$v_{MAP} = argmax \frac{p(a1,a2 \ldots an|v_j\, p(v_j)}{p(a_1,a_2 \ldots\ldots a_n)} \qquad (2)$$

$$v_{NB} = argmax\ P(v_j)\ \ P(a_i \backslash v_j) \qquad (3)$$

where $v_j \varepsilon$ V

Now we could attempt to estimate the two terms in equation 2 based on the training data. It is easy to estimate each of the $p(v_j)$ simply by counting the frequency with which each target value $v_j$occurs in the training data. However, estimating the different $p(a_1, a_2 \ldots . a_n|v_j)$ terms in this fashion is not feasible unless we have a very, very large set of training data. The problem is that the number of these terms is equal to the number of possible instances times the number of possible target values. Therefore, we need to see every instance in the instance space many times in order to obtain reliable estimates.The naive Bayes classifier is based on the simplifying assumption that the attribute values are conditionally independent given the target value. In other words, the assumption is that given the target value of the instance, the probability of observing the conjunction$a_1, a_2 \ldots . a_n$ is just the product of the probabilities for the individual attributes :

$$p(a_1, a_2 \ldots . a_n|v_j) = \prod_i p(a_i|v_j)\ .$$

Substituting this into Equation **(**we have the approach used by the naive Bayes classifier.

Naive Bayes classifier:

$$v_{NB} = argmax\ p\ (v_j)\ \ \prod_i p(a_i\,|v_j) \qquad (4)$$

where v_{NB} denotes the target value output by the naive Bayes classifier. Notice that in a naive Bayes classifier the number of distinct $p(a_i|v_j)$ terms that must be estimated from the training data is just the number of distinct attribute values times the number of distinct target values-a much smaller number than if we were to estimate the $p(a_1, a_2 \ldots . a_n|v_j)$ terms as first contemplated.To summarize, the naive Bayes learning method involves a learning step in which the various and $p(a_i|v_j)$ terms are estimated, based on their frequencies over the training data. The set of these estimates corresponds to the learned hypothesis. This hypothesis is then used to

IJSETT International Journal
"International Journal for Science and Emerging Technologies with Latest Trends"
4(1): 25-24 (2012)
ISSN No. (Online) : 2250-3641, ISSN No. (Print):2277-8136

classify each new instance by applying the rule in Equation .Whenever the naive Bayes assumption of conditional independence is satisfied, this naive Bayes classification v_{NB} is identical to the MAP classification. One interesting difference between the naive Bayes learning method and other learning methods we have considered is that there is no explicit search through the space of possible hypotheses (in this case, the space of possible hypotheses is the space of possible values that can be assigned to the various $p(v_i)$ and $p(a_i|v_j)$ terms. Instead, the hypothesis is formed without searching, simply by counting the frequency of various data combinations within the training.

B. Simulation and Numerical results

The data set was taken by considering the previous year's data. Using visual c# under the naive bayse algorithm analysis classification technique this training set was executed. The data is collected from statistics of the metric data of the car here we select attributes to introduce prediction. For each dataset, we sort the attributes in descending order according to their information gain index and *car* evaluation. In this section, we report our experimental results. Since our algorithm can only process categorical attributes, datasets containing only categorical attributes. In this group of experiment, we show the classification performance when different attributes are selected For each dataset, for various number of attributes are presented.Dataset is taken from these websites-

1) http://www.theautomotiveindia.com
2) http://www.team-bhp.com

Table1: Car Sale Data Month wise(Jan'10-Mar'10)

SN	Company	Jan	Feb	Mar
1	Tata	26245	26985	27761
2	Fiat	2302	2256	2107
3	Toyota	5989	5993	6905
4	Hyundai	29601	31001	31501
5	M & M	14998	12813	14609
6	Skoda	1881	1805	1824
7	Mercedes	403	439	440
8	Maruti	81088	84765	95123
9	Ford	2453	3,223	9478
10	Chevrolet	9421	11111	11330
11	Honda	5983	6275	4338
12	Nissan	40	44	89
13	Audi	306	252	220

Table2: Car Sale Data Month wise(Apr'10-Jun'10)

SN	Company	April	May	June
1	Tata	23102	21324	27811
2	Fiat	1800	2163	2137
3	Toyota	6001	6050	6180
4	Hyundai	28501	27151	27366
5	M & M	12923	13939	13316
6	Skoda	1285	1381	1638
7	Mercedes	296	410	388
8	Maruti	71748	75109	72812

IJSETT International Journal
"International Journal for Science and Emerging Technologies with Latest Trends"
4(1): 25-24 (2012)
ISSN No. (Online) : 2250-3641, ISSN No. (Print):2277-8136

9	Ford	7509	8080	7269
10	Chevrolet	10601	8225	9539
11	Honda	2787	4067	4595
12	Nissan	47	56	77
13	Audi	189	200	233

Table 3: Car Sale Data Month wise(Jul'10-Sep'10)

SN	Company	Jul	Aug	Sept
1	Tata	27,864	25196	23877
2	Fiat	2301	1812	1650
3	Toyota	6834	6361	6235
4	Hyundai	28811	28601	31751
5	M & M	12825	13796	17537
6	Skoda	1222	1512	1467
7	Mercedes	520	573	663
8	Maruti	100857	92674	95148
9	Ford	8,739	7925	8380
10	Chevrolet	7124	7889	8617
11	Honda	4685	5518	7728
12	Nissan	1005	1249	1256
13	Audi	226	273	302

Table 4: Car Sale Data Month wise(Oct'10-Dec'10)

SN	Company	Oct	Nov	Dec
1	Tata	24478	15,340	19706
2	Fiat	2025	1,025	1812
3	Toyota	5650	5,242	7657
4	Hyundai	34725	31,540	26168
5	M & M	16,987	12323	15601
6	Skoda	1694	1841	2617
7	Mercedes	430	518	710
8	Maruti	107555	102,503	89469
9	Ford	9026	7,504	4,301
10	Chevrolet	10051	8364	8468
11	Honda	5275	4105	5135
12	Nissan	1340	1079	1104
13	Audi	357	256	212

Table 5: Car Sale Data Month wise(Jan'11-Mar'11)

SN	Company	Jan	Feb	Mar
1	Tata	30212	31909	27678
2	Fiat	2150	1839	1860
3	Toyota	9185	9308	9726
4	Hyundai	30306	32629	31822
5	Mahindra	17208	14288	17320
6	Volkswagen	5600	7077	8080
7	Nissan	1857	2081	2101
8	Skoda	2825	2512	3009
9	Maruti	10042	101543	110424
10	Ford	10026	9293	10485
11	Chevrolet	969	9283	9365
12	Honda	6358	4843	3576

IJSETT International Journal
"International Journal for Science and Emerging Technologies with Latest Trends"
4(1): 25-24 (2012)
ISSN No. (Online) : 2250-3641, ISSN No. (Print):2277-8136

13	Audi	480	450	681

Table 6: Car Sale Data Month wise(Apr'11-Jun'11)

SN	Company	April	May	June
1	Tata	2338	19401	21994
2	Fiat	2030	2143	1800
3	Toyota	9680	7470	12032
4	Hyundai	31636	31123	30402
5	Mahindra	15459	16702	16053
6	Volkswag	6995	6177	5395
7	Nissan	1216	1588	1632
8	Skoda	2446	2761	2611
9	Maruti	87144	93519	70020
10	Ford	7319	7046	7023
11	Chevrolet	10021	8292	8187
12	Honda	2012	2334	3455
13	Audi	375	408	408

Table 7: Car Sale Data Month wise(Jul'11-Sep'11)

SN	Company	Jul	Aug	Sept
1	Tata	17192	16829	26319
2	fiat	1100	1067	818
3	Toyota	13592	11679	12807
4	Hyundai	25642	26677	35955
5	Mahindra	17312	15664	19447
6	Volkswagen	6528	6089	6847
7	Nissan	1593	1384	2176
8	Skoda	2411	1812	1654
9	Maruti	66504	77086	78816
10	ford	7587	7382	7801
11	Chevrolet	9465	9012	10078
12	Honda	4725	6907	4758
13	Audi	343	510	555

Table8: Car Sale Data Month wise(Oct'11-Dec'11)

SN	Company	Oct	Nov	Dec
1	Tata	25124	27,737	28916
2	fiat	622	1,036	500
3	Toyota	10762	13,956	15948
4	Hyundai	33001	35,000	29516
5	Mahindra	18,756	17813	19341
6	Volkswagen	7266	6722	5598
7	Nissan	2990	2,688	1596
8	Skoda	2134	2627	3938
9	Maruti	51,458	82870	77475
10	Ford	8091	8322	5979
11	Chevrolet	10009	8382	8993
12	Honda	5526	1982	1072

IJSETT International Journal
"International Journal for Science and Emerging Technologies with Latest Trends"
4(1): 25-24 (2012)
ISSN No. (Online) : 2250-3641, ISSN No. (Print):2277-8136

13	Audi	485	479	467

Results:-

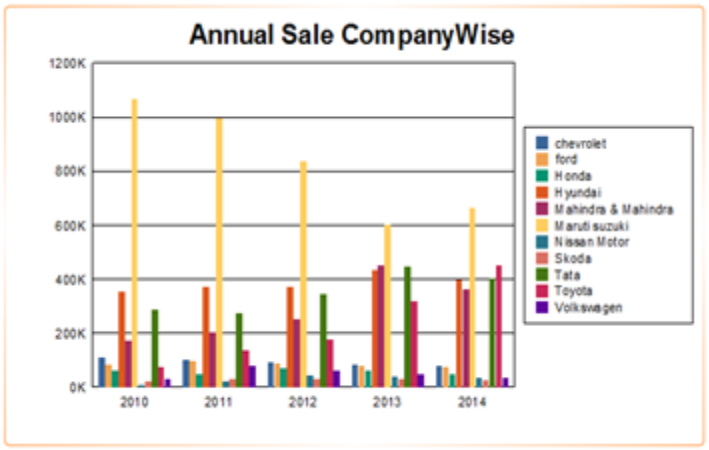

Figure 2: Annual sale company wise(prediction)

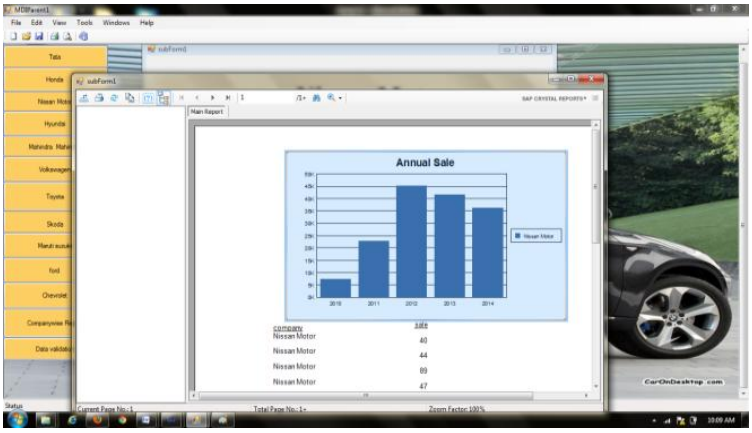

Figure 3: Annual sale of Nissan(prediction)

IJSETT International Journal
"International Journal for Science and Emerging Technologies with Latest Trends"
4(1): 25-24 (2012)
ISSN No. (Online) : 2250-3641, ISSN No. (Print):2277-8136

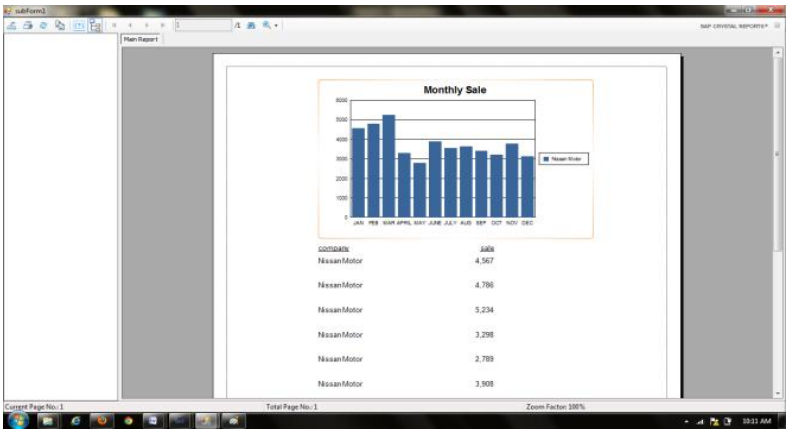

Figure 4: Predict monthly sale Nissan(2012)

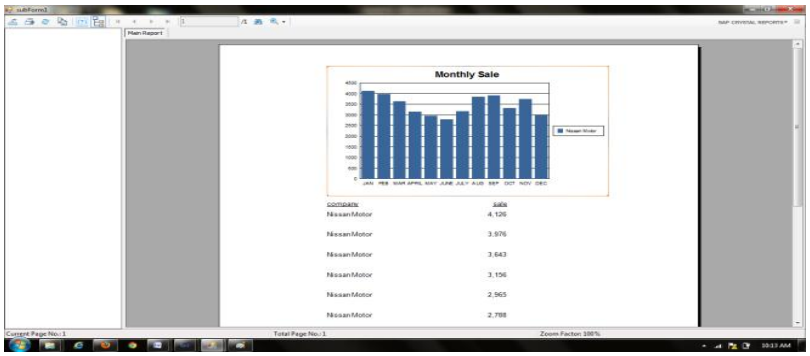

Figure 5: Predict monthly sale Nissan(2013)

Figure 6: Predict monthly sale Nissan(2014)

IV.CONCLUSION AND FUTURE WORK

This paper introduces the application of data mining technology in the car manufacturing unit and obtains an analysis result from large data and predicts the future of number of cars with the help of naïve bayse algorithm. Naïve bayes algorithm is a popular data mining algorithm which can be used in a number of application areas. Four predictor variables are used in this prediction for getting the accurate result. The results are validated by comparing them with the real values.

The future scope of this work can be that Naïve bayse algorithm can be implemented using more predictor variables rather than four predictors. More predictor variables mean more accuracy. Further work is under progress to develop an algorithm which can handle multiple predictor variables.
Also, the future can be predicted for a larger number of years in advance so that car manufacturing industries can have more profit out of this application.

References

[1] Dr. M.Hanumanthappa1, Sarakutty.T.K "Predicting the Future of Car Manufacturing Industry using Data Mining Techniques",Feb ,2011.

[2] Abhijeet Kamble, Ajay Nalawade, Ashutosh Deo,Sagar Ranadive, "Demand/Sales Forecasting in Indian Firms", April 3,2006.

[3] Driving Strategic Planning with Predictive Modeling: An Oracle White Paper.

[4] Alex Berson, Stephen Smith, and Kurt Thearling "Building Data Mining Applications for CRM" .

[5] Mouhib Al-Noukari, Arab International University, Damascus,Syria & Wael Al-Hussan, The Arab Academy for Banking and Financial Sciences,Damascus, Syria, "Using Data Mining Techniques for Predicting Future Car market Demand : DCX Case Study".

[6] Jiawei Han & Micheline Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers.

[7] David L. Iverson, "Data Mining Applications for Space Mission Operations System Health Monitoring", NASA Ames Research Center, Moffett Field, California, 94035, 2008.

[8] Ping Lu, Brent M. Phares, Terry J. Wipf and Justin D. Doornink,"A Bridge Structural Health Monitoring and Data Mining System",in Proceedings of the 2007 Mid-Continent Transportation Research
Symposium, Ames, Iowa, August 2007.

[9] Tibebe Beshah Tesema, Ajith Abraham And Crina Grosan, "Rule Mining And Classification of Road Traffic Accidents Using Adaptive Regression Trees", In Proc. Of I. J. On Simulation, Vol. 6,No. 10 and 11, 2008.

[10] Chitriki Thotappa and Dr. Karnam Ravindranath " Data mining Aided Proficient approach for optimal inventory [1]control in supply chain management" (IJCSIS) International Journal of Computer Science and Information Security, Vol. 8, No. 2, May 2010

[11] S.Shankar, T.Purusothaman, "Utility Sentient Frequent Itemset Mining and Association Rule Mining: A Literature Survey and Comparative Study", International Journal of Soft Computing
Applications, ISSN: 1453-2277 Issue 4 (2009), pp.81-95.