**Increase overall sales of the company on that thesis find the growth of company for next one year**

**Internship Report**

**Submitted by**

RITU KUMARI

Roll No: COE15B032

in partial fulﬁllment for the award of the degree of

Bachelor of Technology

In

Computer Engineering



Indian Institute of Information Technology

Design and Manufacturing, Kancheepuram, India

Problem Definition

How to increase overall sales on that thesis find the growth of company for next one year.

There is company of e commerce want to predict data for in

There is an E-Commerce company that wants to analyze their data in order to grow its business, they want to see their company growth of next one year because they want to prepair its storage and increase the Business.

2

Acknowledgement

For the ancestor who proved the path before me upon whose shoulders I stand. This is also dedicated to my family and friends who supported me on this journey.

I would like to express my deepest gratitude to my adviser Mr. Anup Kumar for his guidance, assistance and support throughout this project. His continuous inputs, comments, remarks and engagement through the learning process of internship work are highly invaluable to me.

Finally, I would like to Thanks FuturoTech Prt Ltd for providing this work and they had trust on me for this project

3

|  |  |
| --- | --- |
|  | |
|  |
|  |
|  |

Contents

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 [ABSTRACT](#page1) | | | | 5 |
| [2](#page5) | [Introduction](#page5) | | | 6 |
| 3 | [Preliminaries](#page8) | | | 8 |
| [4](#page9) | [Our Contribution](#page9) | | | 9 |
|  |
|  |
|  |
|  |
| [5](#page21) | Hypothesis generation | | | 14 |
|  | [5.1](#page21) | | [Create new attribute](#page21) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 15 |
|  | [5.2](#page24) | | Model setup on the data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . | 16 |
| [6](#page25) | Over all result | | | 18 |
|  | |
| [8](#page30) | [Future Work](#page30) | | | 20 |

4

**ABSTRACT**

Sales prediction is a must famous problem in the industry, where company wants to know about their sales that how to increase the over all sales and after few years what will be the condition of the company. I got one of the sales prediction problem for my internship, this company was e commerce company where I had to find “how to increase sales of company on that basis what will be condition of the company in next one year”. There was data about store, sales record, product feature, cost of the product, area of store, brand of the product etc. This data can put to proper use to predict the result of sales that’s why company wants to solve this problem.

This report aims at solving the problem of predict the result of sales by identifying the important attributes from the data set using data mining algorithms. I saw previus work technique of the company and used as a reference were want to increases over all sales. My report describe in detail the different attributes selection technique as well as the data mining algorithms use to solve this given problem of company.

I have also used accuracy as the evaluation criteria to evaluate how to well prediction performance. Some of future work is also suggest in this report.

My work witch used to predict the result of the increasing sales of the company was built successfully with accuracy rate of about 60% to 89%. The list of attributes was cut down to 10 important out of the 21 attributes variable in the data set by using the attribute selection algorithm. The 10 data mining algorithm that were performed on the model like Decision Trees, Error Metrics, Random Forest , Linear Regression etc.

Data mining algorithm.

Key word- Data mining, prediction, sales

5

**Introduction**

This report is a short description of my 5 month internship carried out as compalsury concept of the B Tech.

Sales growth is a most common problem for every company. I work for FuturoTech where had to find increase overall sales rate and on that thesis what will be the condition of company in next one year.

**Sales Forecasting** is the process of estimating what business’s sales are going to be in the future. A sales forecast period can be monthly, quarterly, half-annually, or annually.

Sale forecasting is an integral part of business management. The purpose of sales forecasting is to provide information that we can use to make intelligent business decisions.

Sales forecasting is the process of estimating future sales. Accurate sales forecasts enable companies to make informed business decisions and predict short-term and long-term performance. Companies can base their forecasts on past sales data, industry-wide comparisons, and economic trends.

Sales forecasting gives insight into how a company should manage its workforce, cash flow, and resources. In addition to helping a company allocate its internal resources effectively, predictive sales data is important for businesses when looking to acquire investment capital.

Sales forecasting allows companies to:

* Predict achievable sales revenue;
* Efficiently allocate resources;
* Plan for future growth.

Salas prediction is categorized in forecasting problem in data mining. It increases with time (we can also say time series problem).

6

Forecasting-: **Forecasting** is determining what is going to happen in the future by analyzing what happened in the past and what is going on now. It is a planning tool that helps business people in their attempts to cope with the uncertainty of what will might and might not occur. Forecasting relies on past and current data and analysis of trends. Company management, government departments, economists, and investors utilize forecasting to decide how to allocate their resources and prepare reports. They also use it to plan for anticipated expenses.

As far as companies are concerned, this is mainly based on predicted demand for its goods or services.

Economists, for example, might estimate some variable of interest rates at a specific date in the future. Forecasting is similar to **prediction**.

**Time series analysis-:**

Time series analysis is a that statistical technique that deals with time series data, or trend analysis.  Time series data means that data is in a series of particular time periods or intervals.  The data is considered in three types:

**Time series data:**A set of observations on the values that a variable takes at different times.

**Cross-sectional data:** Data of one or more variables, collected at the same point in time.

**Pooled data:** A combination of time series data and cross-sectional data.

Moreover, time series analysis can be classified as:

* 1. Parametric and Non-parametric
* 2. Linear and Non-linear and
* 3. Univariate and multivariate

7

**Preliminaries**

Company sale’s home need things like shop, bathroom wash, room cleaner etc. It has 10 store in different city of India.

8

**Our Contribution**

Data cleaning

I got raw data from the company and their was information about product cost, store, quility, brand etc. so firstly I had to clean the data for better result.

In the cleaning process 1st I worked on missing value in the data. There were many type of data so I had to choose suitable process for different attributes

Missing data can have a severe impact on building predictive models because the missing values might be contain some vital information which could help in making better predictions.

**Method to fill missing value**

For choose best method to missing value I created small subset of the given data. From that data, deleted manually few value and try to impute that value with multiple method like KNN imputation, central statistics and Prediction method, then compare that all the predicted value with the actual value and see which one is closer, then choose that method as best method.

**I use statistical method**

**Deletion of rows**: In train dataset, observations having missing values in any variable are deleted. The downside of this method is the loss of information and drop in prediction power of model.

**Mean/Median/Mode Imputation**: In case of continuous variable, missing values can be replaced with mean or median of all known values of that variable. For categorical variables, we can use mode of the given values to replace the missing values.

**9**

**central statistics**

For continuous variable

1.Mean

2.Median

For categorical variables

3.Mode Imputation

1. **Building Prediction Model**: We can even make a predictive model to impute missing data in a variable. Here we will treat the variable having missing data as the target variable and the other variables as predictors. We will divide our data into 2 datasets—one without any missing value for that variable and the other with missing values for that variable. The former set would be used as training set to build the predictive model and it would then be applied to the latter set to predict the missing values.

10

2. Hypothesis generation

It is like mind map where can I pur our ideas about data.

Type of costomer

Area of store

Sales can affect sales growth

Environment product type

11

**Create new attributes and drop some attributes**

->I think the given attributes in a dataset are not sufficient to give satisfactory predictions. So that I created new features which might help in improving the model’s performance.

**->Need to change data type**

I changed data type of the data set because some of the algorithm fit with some particular data type like I changed categorical to numerical

-**>**I choose data scaling method (Normalization/Standardization)

This stuff was very important when dealing with the parameters of different units and skills.

**->**I counted the frequency of occurrence of the same phenomena in two different population with different size with different unit and want to compare them, I had to normalize both, because otherwise I do not know how big the influence of our phenomena is in relation to the total number of cases. Thus, data scaling is needed, when comparing populations/phenomena of different size but with the same origin.

->Want to drop some attributes

For droping some attributes I used chi-square test that is because some of the attributes are same work so no need to carry both of them, I have to choose one of attributes.

12

**chi-square test-:** Chi-squared is a means of statistically evaluating data. It is used when categorical data from a sampling are being compared to expected or "true" results

**Calculate**

Determine the critical p value that we will use to evaluate our data. This is the percent probability (divided by 100) that a specific chi-square value was obtained by chance alone. Another way of thinking about p is that it is the probability that our observed results deviated from the expected results by the amount that they did solely due to random variation in the sampling process.

Look up the p value associated with our chi-square test statistic using the chi-square distribution table. To do this, look along the row corresponding to our calculated degrees of freedom.

Find the value in this row closest to our test statistic. Follow the column that contains that value upwards to the top row and read off the p value. If my test statistic is in between two values in the initial row, I can read off an approximate p value intermediate between two p values in the top row.

* I get low value for chi-square means there was a high correlation between our two sets of data

13

Model setup on the data

**Hypothesis testing with these algorithms**

1. Decision Trees
2. Error Metrics
3. Random Forest
4. Linear Regression
5. Logistic Regression
6. Visualizations

7. Time series Analysis

1. XGBoost

2.ARIMA models

3. Box-Jenkins multivariate models

14

**Decesion tree**

I used Decision Trees for select my attributes. I had branching type of attribute like witch type of family will buy more

• IG = Entropy of the system before split – Entropy of the system after split

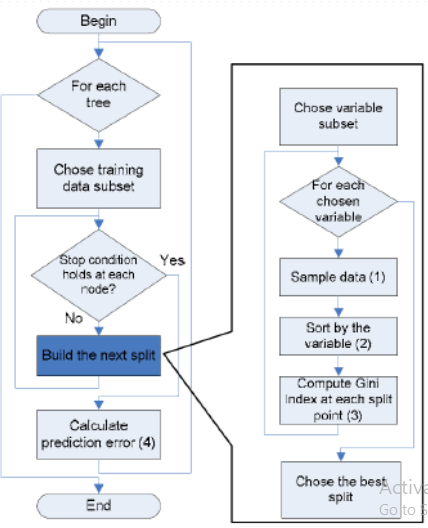
• Entropy: Uncertainty in the data/Measure of impurity

Selects the variable whose Information gain (GI)is high

15

Algorithm Flow chart using random forest

* I used Random Forest for my categorical data like example
* X1 type of family in y1 type of session from Z1 type of store
* Will buy p1 type of item or not



16

**ARIMA models**

I used this model for time series analysis where I had to find company condition in next one year.

**How did I used it**

It has three ordered parameters (p,d,q).

* p is the order of the autoregressive model(number of time lags)
* d is the degree of differencing(number of times the data have had past values subtracted)
* q is the order of moving average model

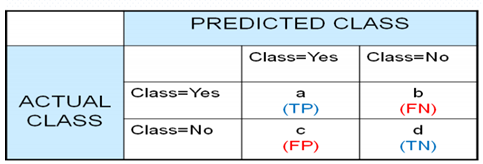
**Box-Jenkins multivariate models**

* I had to forecast the value for next one year that is time series.
* Their was seasonality significant so I used this model
* The first step in developing a Box–Jenkins

17

**Error Metrics**

For check my model accuracy I used error matrix for each model



Accuracy = TP+TN/ Total observations

**Conclusion of 1st part**

Here I found best results from random forest there is sales increase 65% to 89%.

**2nd part**

Problem-> find the company condition in next one year

For this part of problem I used ARIMA models and Box-Jenkins multivariate models

18

**Find out the condition in next one year of the company, basis of predicted data**.

After got value of predicted data that increased by 65% to 89% I had to find out forecasting value.

For this problem I use linear regression algorithm as well as Time series analysis.

Time series analysis gave better result than linear regression.

From unpredicted data, company was increasing only 12% sales in next one year but according to predicted data it will be increase 36.446%.

19

**Future work**

I used existing algorithm for this particular data and get 89% accuracy. Someone who want to do work in future then they can create different hypothesis on that basis they can also create new algorithm. I got 89% accuracy they can improve it more than this.

20