Abstract:

The aim of our work is to compare different type of predictive modeling over the transactional dataset of superstore business. We've planned to implement some recent and classic data mining, machine learning and statistical modeling algorithms to make specific decision based on multiple features of the dataset and then compare the accuracy and performance ratio of different algorithms for the specific prediction.

Introduction:

Modern business process generates data for each transaction. Within a few transactions, a business process makes lot of data which is valuable. Many interesting patterns, solution and insights can be found by studying those data properly. It's also possible to determine the approximate revenue of next year by analyzing previous business data. The field of data analysis performs such determinations by studying given data is generally known as Predictive analysis.

Literature review:

Predictive Analytics [1]:

Predictive analytics a branch of advance analytics that is used to make prediction about unknown future events. It uses many techniques from machine learning, data mining, statistic modeling and artificial intelligence to analyze data to create prediction about future events. It uses several data mining, machine learning, predictive modeling and analytical techniques to bring together the management, information technology and modeling business process to make prediction about future. The pattern found in historical and transactional data can be used to identify risks and opportunities for future. By Successfully applying predictive analytics the business can effectively interpret their data for future benefit.

Books and Articles review:

To understand the basic concepts of data mining, machine learning and predictive modeling we've studied several sections of different books, articles. Data Mining Concept and Techniques [i] written by Jiawei Han and Micheline Kamber is one of them where the process and steps of data mining were described. We've gathered the idea predictive analytic from this book and gone through some examples. Machine Learning A Probabilistic Approach [2] written by Kevin P Murphy (one of the google researchers) is another book we have studied for the mathematical and conceptual basics of the algorithms. Moreover we've studied two journals regarding predictive analysis and modeling in different fields to understand the scopes and opportunities of the topic.

Dataset for Predictive analysis:

We have selected an opensource dataset to perform the task. The dataset contains the information about global superstore transactions between the year 2011 to 2015. Full description of the dataset is given as the appendix of this proposal.

Tools and Platforms:

We have selected Python as the programming language for the analysis and Pandas for handling the dataset. Scipy, Numpy, Scikit Learn and other python based libraries for statistical, predictive and machine learning operation. Moreover, currently available or opensource IDEs/editor and platforms will be used to complete the task.

Methodology:

We divided the whole process into these sub processes to reach the final target. The sub processes are:

- Making the question based on what we really want to predict
- Data Collecting
- Data preprocessing
- Data Model Creation
- Apply algorithms to make predictions based on different features.
- Visualizing the predictions.
- Comparison graph for different approaches.

Data Collecting:

We have already collected the targeted dataset for the task. The dataset is **Opensource** and collected from **Tableau Community** [2]**.**

Making the question based on what we really want to predict:

The first and foremost goal of our task is to make the right question because based on our question we will predict the answer. The question must be meaningful and relevant to make a fully functional predictive model. Question may contain a single line but in order to achieve our goal we have to analyze the question more thoroughly. Analyzing the question means find out the statement beneath the question. Another thing should remember that to answer the question we must have necessary data. Making the right question is the foundation of our task. We also must select our algorithm according to the question.

Data preprocessing:

In this step, the dataset will be checked thoroughly and removed the unnecessary attributes or missing values. Only the targeted attributes will be kept for the modeling and further use. After this step, the dataset will be clean and customized for our next steps.

Data Model Creation:

This is the third step of our whole process and one of important processes for the final goal. Here we will create different data models for different prediction algorithms by using libraries described in the tools and platforms section.

Make Prediction:

This is the main stage of our total work. We will different algorithms to make predictions. As our dataset is supervised (labeled data) we have decided to apply these algorithms for prediction.

1.Support Vector Machine(SVM) [3] :

Support vector machine (SVM) learns a hyperplane to classify data into 2 classes. At a high-level, SVM performs a similar task like C4.5 except SVM doesn’t use decision trees at all. This is a supervised learning, since a dataset is used to first teach the SVM about the classes. Only then is the SVM capable of classifying new data. More specifically a support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

The margin is the distance between the hyperplane and the 2 closest data points from each respective class. The key is:

SVM attempts to maximize the margin, so that the hyperplane is just as far away from 2 classes. In this way, it decreases the chance of misclassification.

2.Apriori [4]:

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. Apriori is generally considered an unsupervised learning approach, since it's often used to discover or mine for interesting patterns and relationships. Association rule learning is a data mining technique for learning correlations and relations among variables in a database. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis. So, this algorithm is one of the most important topic in project because it is designed to operate on databases containing transactions. There are 3 important concepts to understand Apriori algorithm:

1. The first is the size of itemset. Find those items that tend to be purchased together more frequently than other items. The ultimate goal being to get shoppers to buy more. Together, these items are called item sets.
2. The second is support or the number of transactions containing the itemset divided by the total number of transactions. An itemset that meets the support is called a **frequent** itemset.
3. The third is confidence or the **conditional probability** of some item given that have certain other items in the itemset.

The basic Apriori algorithm is a 3-step approach:

1. **Join:** Scan the whole database for how frequent 1-itemsets are.

2. **Prune:** Those item sets that satisfy the **support** and **confidence** move onto the next round for 2-itemsets.
3. **Repeat:** This is repeated for each itemset level until we reach our previously defined **size**.

Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. It generates candidate item sets of length ***k*** from item sets of length ***k-1***. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent ***k***-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates.

3.Polynomial Regression [5]:

Polynomial regression is a form of regression analysis in which the relationship between the independent variable ***x*** and the dependent variable ***y*** is modelled as an *n*th degree polynomial in ***x***. Polynomial regression fits a nonlinear relationship between the value of *x* and the corresponding conditional mean of ***y***, denoted E(*y* |*x*).

The goal of regression analysis is to model the expected value of a dependent variable *y* in terms of the value of an independent variable (or vector of independent variables) *x*. In simple linear regression, the model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

is used, where ε is an unobserved random error with mean zero conditioned on a scalar variable *x*. In this model, for each unit increase in the value of *x*, the conditional expectation of *y* increases by β_1 units.

In general, we can model the expected value of *y* as an *n*th degree polynomial, yielding the general polynomial regression model

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon$$

4.C4.5 [6]:

C4.5 constructs a classifier in the form of a decision tree. In order to do this, C4.5 is given a set of data representing things that are already classified. Since the training dataset is labeled with classes, this is supervised learning. The decision trees generated by C4.5 can be used for classification. C4.5 builds decision trees from a set of training data in using the concept of information entropy. The training data is a set ***S=s₁, s₂…*** of already classified samples. Each sample s_i consists of a p-dimensional vector $(x_{1,i},x_{2,i},\ldots\ldots, x_{p,i}\ )$ where the x_j represent attribute values or features of the sample, as well as the class in which s_i falls. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub lists.

Concrete Plan:

The methodologies are described above are the general work flow of our task. In a structured manner, the summery and concrete plan is given here:

- **Collect and organize the relevant resources:** The very first step of our task is to collect the information and tools and prepare initial documentation.
- **Analyze the Documentation:** The chosen dataset and documentation of collected resources will be analyzed to find out the possible questionnaires which will be used to select the right predictive algorithm for the specific task.
- **Background Study:** Primary prerequisite of applying an algorithm is the background study. We will make document and study the mentioned algorithms and other requirements so that the implementation phase becomes smoother.
- **Test and Implementation of Each Algorithm:** We will implement the algorithm (at least 3 of the mentioned) to make predictive model and calculate the performance and accuracy of each algorithm. By the time we will document this information into the final documentation.
- **Merging the Result:** Individual results the algorithms will be merged and compared to find the fittest model. A comparison graph will also be created and documented.
- **Final Documentation:** After completing these steps final documentation will be written and submitted.

Tasks we've Completed so far:

We have started our planning from the beginning of the semester and gathered different ideas based on our capabilities and interest. Then we've sorted the ideas and finally selected the goal. As soon as the goal is selected we've started to study the background of the topic and came up with the idea of **Predictive Analytics.**

- Then we searched for the proper dataset and found one interesting to us. With the consent of our honorable supervisor, we selected it as final.

- We analyzed the dataset to understand the meaning of the features and documented it.
- Completed the initial guesses of prediction algorithms.
- Studied the overviews of the algorithms.

Conclusion:

We make prediction everywhere in our life, from selecting daily needs to scientific research. In this data driven world we generate valuable data in every moment which can be used as a landmark to guess the future events. Modern business is also a transactional data driven process where millions of data is generated every day. We’ve taken a small fraction of data of such transactions to predict the future or find specific answer from data. To complete the task, we’ve mentioned the strategies and steps in this proposal. By applying these steps, one can easily make predictive model and compare them with other models to find the best result.

References

I. Jiawei Han, M. K. (2011). *Data Mining: Concepts And Techniques* (Vol. III). USA: Morgan Kaufmann.

II. Murphy, K. P. (2012). *Machine Learning a Probabilistic Approach.* Cambridge: The MIT Press.

III. S.B. Soumya, N. D. (2016). Data Mining With Predictive Analytics for Financial Applications . *International Journal of Scientific Engineering and Applied Science, 2*(1), 8.

IV. Usama Fayyad, G. P.-S. (1996). KDD-96: Proceedings : Second International Conference on Knowledge Discovery & Data Mining. *ASSOCIATION FOR THE ADVANCEMENT OF ARTIFICIAL INTELLIGENCE*, 391.

Few website links:

[1] : https://www.predictiveanalyticstoday.com/what-is-predictive-analytics/

[2] : https://www.tableau.com/sites/default/files/training/global_superstore.zip

[3] : http://www.kdnuggets.com/2015/05/top-10-data-mining-algorithms-explained.html

[4] : http://www.kdnuggets.com/2015/05/top-10-data-mining-algorithms-explained.html

[5] : https://en.wikipedia.org/wiki/Polynomial_regression

[6] : http://www.kdnuggets.com/2015/05/top-10-data-mining-algorithms-explained.html