DESIGN AND IMPLEMENTATION OF PREDICTIVE MODELING FOR IRIS,CHURN AND WEATHER DATA

Nikshep A[1], Rakesh C R[2], Dheeraj B[3] , Darshan D[4] , Sudeep J[5]

1. Student, Information Science & Engineering, NIE Institute of Technology, Karnataka, India
2. Student, Information Science & Engineering, NIE Institute of Technology, Karnataka, India
3. Student, Information Science & Engineering, NIE Institute of Technology, Karnataka, India
4. Student, Information Science & Engineering, NIE Institute of Technology, Karnataka, India
5. Assistant professor, Information Science & Engineering, NIE Institute of Technology, Karnataka, India

ABSTRACT

The objective is to provide the data analysts or data science users to give much more powerful tool which would help them understand the patterns of data in better way. The tool would analyze the data and provide details such as outliers, imbalanced fields, missing values, performance of individual field and much more. This would help the data analysts to skip the manual process of cleaning the data before using the data to build a predictive model. Once the data is analyzed and cleansed, based on the business objective one can choose the outcome field that whose value should be predicted and then the tool would show the correlation between the different fields in the data set. This would help the data analyst to remove the redundant fields from being used in the model building. Once the data analyst finalizes the fields to be used, a predictive model (Regression or Decision Tree) would be built from the tool automatically using JRI(Java/R Interface) that visualizes the decision tree that would be built for the dataset and would be available for download in PMML format (http://dmg.org/pmml/v4-3/GeneralStructure.html). One can download the model and use it to execute using any third party tools that are available in market like JPMML and other scoring engines. This would help the data analysts save time by removing the major manual task of data cleaning and model building. This tool also makes sure that it could be used by no-vice users as well.

Keywords:- *Predictive model, Iris dataset, Churn dataset, Weather dataset, Decision tree, PMML.*

1. INTRODUCTION

Though many companies have adopted Business Intelligence solutions and it enables slicing and dicing of their data and provides detailed view of what's going on, they are challenged on getting insights into the future ("What should we do") or ("What will happen next"). The lack of predictive analytical capabilities hinders organizations to take the right decision at the right time. The goal of this project is to provide the detailed insights in to the data that will be used for predictive modeling and help to make the data better (in terms of cleansing) and use the cleansed data for building the predictive models to gain higher accuracy and higher performance for any kind of data. Predictive analytics comprises of a variety of techniques from statistics and data mining that analyze current and historical facts to make predictions about future events. In business, the predictive model exploits hidden patterns found in historical and transactional data and predicts the probable future outcomes with a certain degree of accuracy. These Predictive models capture relationships among many factors in the transactional data associated with a particular set of conditions, guiding decision making for candidate transactions. Basically these models ensure that the actions taken today will directly achieve the organization's goals tomorrow. That's the way a Predictive analytics model works, and that's what gives it competitive advantage in the marketplace. Predictive analytics can help companies optimize existing processes, better understand customer behavior, identify unexpected opportunities and anticipate problems before they happen.

1.1 Data Analysis

The data can be cleansed in terms of the following and certain characteristics of this dataset can be computed:

1. Outliers
2. Skewness
3. Imbalanced
4. Correlation

The above mentioned characteristics are a part of the process of ***data cleaning,*** where the unwanted and irrelated data are removed from the dataset in order to have a good predictive power for the complete dataset.

1.1.1 Outliers

Outliers are handled in a careful manner by the method of ***IQR (Inter Quartile Range)*** where the complete dataset is divided into three quartiles and in each quartile, the outliers are found by the standard formula of **IQR.**

1.1.2 Skewness

Skewness will be handled by the efficient method of ***Kurtosis.*** This method will classify the data points on the graph Such that there are no negative or positive skewness.

1.1.3 Imbalanced

Imbalanced is the problem that occur for symbolic fields in the dataset. This problem will be handled by using the simple regular expressions to classify numeric fields from the symbolic fields. This would be able to classify to the imbalance problem.

1.1.4 Correlation

Correlation will be established by the ***Pearson's formula.*** This characteristic feature would be able to build a relation between two variables (if any).

2. ARCHITECTURE DIAGRAM

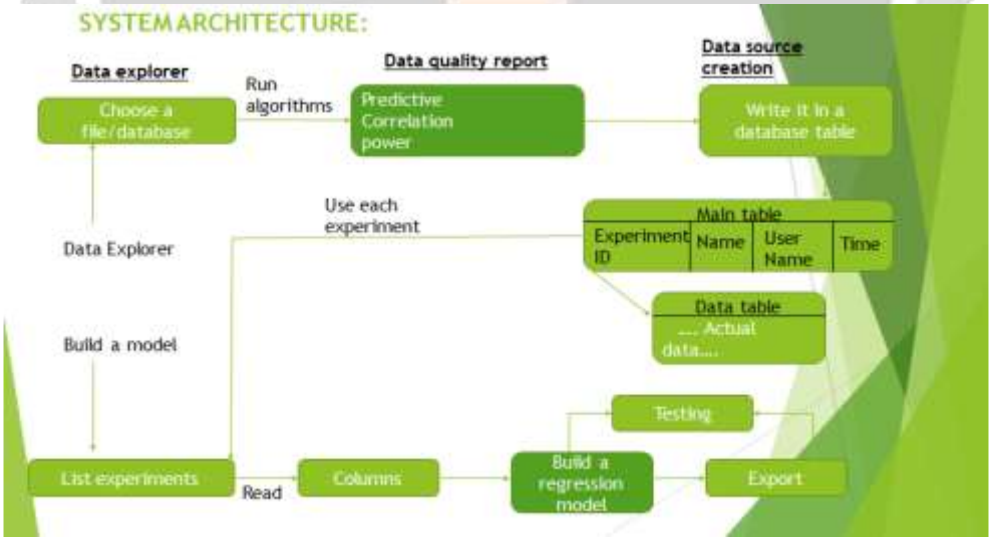

Fig -1:Architecture Diagram

The implementation is divided into two modules:

1. Data Explorer is the module that provides the most use for data analysts and scientists. This module covers the complete analysis of the data set which includes, csv reader, data cleaning and producing the data quality report.

2. Predictive model is the module where the prediction of a target variable is done using the inputs from the user. The predictors will be selected according to their predictive power.

In the first module, the data set will be analyzed using the common characteristics of the data set like outliers, skewness, imbalanced and correlation. All these measures would be used to clean the data and to calculate the predictive power of each field in the data set.

Whereas in the second module, the predictive model would be built for each of the data sets separately and the model would be based on Decision tree/Regression . The tree would also be visualized in the R- studio and the model would be exported in the standard PMML format which can be easily run as an application.

3.Existing system:

Traditional Predictive Analytics

Limitations:

- Each data point must be planned and collected intentionally
- Expensive to design, collect the data

Applications

- Research and development in agriculture and industry
- Designed experiments for product and process improvement
- Verifying the effectiveness of healthcare treatments
- Clinical trials (randomized double blind are best)
- Predicting the weather
- Polls: opinions, election politics, Nielson ratings.
- Standardized tests: e.g. SAT's to predict college success
- Actuarial Science: life expectancy, etc.

4.Proposed system :

- The proposed system will ease the job of a data scientist or business analyst from preventing them to go through massive amount of data manually and do calculations.
- The robust algorithms will provide more accuracy which will reduce the error margin in solving the business problem.
- The data scientist or the business analyst will get more in depth insight into the data which also sorts the predictors based on their predictive power.
- Most importantly, model can be exported in just one click in the standard PMML format and can be executed with any PMML execution engine.

5. CONCLUSIONS

The project satisfies the proposed system by providing the complete data quality report for the data set. The cleansed data set is then used for prediction of the required field. The prediction will be done by considering the inputs from the user and finally the model is exported as a PMML file.

6. REFERENCES

[1]. Analysis of Research in Healthcare Data Analytics : Mohammad Ahmad Alkhatib, Amir Talaei-Khoei, Amir Hossein Ghapanchi

[2]. A REVIEW ON PREDICTIVE ANALYTICS IN DATA MINING : Kavya. V , Arumugam.S

[3]. Data Mining Engine using Predictive Analytics : Sakshi Rungta, Vanita Jain, Akanksha Utreja

[4]. International Journal for Research in Applied Science & Engineering Technology (IJRASET) ©IJRASET: Predictive Analysis of Diseases: An Overview BY Purvashi Mahajan , Abhishek Sharma