5/3/2017

Krishanthan Krishnamoorthy, Sriharsha Bopparaju and Uday Midha

**University of Wyoming**

**Department of Computer Science**

**COSC 4570/5010.**

**Data Mining-2017-G3**

**Determining External Factors Responsible for High Risk Cashiers in A Safeway Store**

**ABSTRACT**

In this project, we intend to determine external factors that may control the risk factor  
of the cashier in the Safeway store. The given Safeway cashier risk report identifies the risky cashiers based on the factors such as total number of transactions, coupon usages, refunds, item voids count, base average bag size etc. And here we try to identify external factors (if any) which may determine or identify the risky cashier. We observed the Safeway cashier report carefully and then it was assessed that few stores have high number of risky cashiers. We considered and researched the external factors such as the population of the city, house hold income of the city and the cashier’s salary, which may influence or control the risk factor of the cashier. We collected the city data by the web scraping, cashier’s salary using the Indeed API and analyzed the collected data using the decision tree. This would be helpful for Safeway to identify high risk stores (stores having more number of high risk cashiers) depending upon the above mentioned external factors.

**1) INTRODUCTION**

Determining External Factors Responsible for High Risk Cashiers in A Safeway Store Project considers the external factors such as population of the city, household income of the city etc. which may drive or control the risk factor of a cashier in a Safeway store. The given Safeway Risk Factor Report [1] identifies the risky cashier based on many factors such as Avg. basket size, Total no. of transactions, Total items count, Coupons etc. And we thought that there may be other factors (such population, household income) which may control the risk factor of a cashier in the Safeway store. In this way, it would be helpful for the Safeway to identify high risky stores. Here in this project we implemented two methodologies such as Web Scraping and Decision Tree. Using the Web Scraping technique, we have collected collect the data such as the Store Numbers and their addresses, Population of the city where the store is located, household income of that city, Avg. Cahier salary and then analyzed the above collected data using the decision trees. The methodologies we used, our research approach and the results are described following the report.

**2) RELATED WORK**

Safeway has already calculated the risk factor of the cashier and identified few risky stores based on Avg. basket size, Total no. of transactions, Total items count, Coupons etc. So in our project we tried to determine the external factors (such as population of the city, household income, cashier’s salary, cashier’s age, cashier’s experience, cashier’s training score (if any)) responsible for high risk cashiers. We considered this would be helpful for the Safeway to identify its risky stores better.

**3) METHODLOGY**

**3.1 Web Scraping**

Web Scraping is a computer software technique of extracting information from websites. Web scraping programming uses World Wide (WWW), and the technique mostly focuses on the transformation of unstructured data (HTML format) on the web into structured data (database or spreadsheet) [6]. Moreover, the Hyper Text Transfer Protocol (HTTP), used to extract the data from websites or through a web program. Even though we can do web scraping by user, the term generally implies automated methodology completed using a bot or web crawler. The gathered and copied data in local database or spreadsheet later used to recuperation or examination.

The popular web scrapping is done in python (we used python in our project for web scrapping). In Python, we can use BeautifulSoup [5], a python library as a tool and we use Urllib2 as an additional major tool to scrapping the data.

1) Urllib2 is a Python module that can used to fetching URLs. The classes and function helps to accomplish the URL action. For example, basic and advanced authentication, redirection, cookies, etc.

2) BeautifulSoup is another nice tool for extracting data from a web page. This tool used to extract any information in the websites, for example tables, lists, paragraph and it can be accomplish by some technique that can apply to HTML tags. Using the HTML tags we can access anything in the websites and we can filter information. (In our project, we have looked at the source code of the HTML and applied the correct technique for extract the information by using the HTML tags.)

Most of the time we need to use Urllib2 and BeautifulSoup together for extracting the data from the websites.

**3.2 Decision Tree**

The Decision Tree is a schematic, tree-shaped diagram which provides a modelling technique that shows a statistical probability, which is easy for humans to comprehend and it simplifies the classification process [7]. A decision tree is a hierarchical structure of nodes and directed edges. Decision trees give people an effective and easy way to understand the potential options of a decision and its range of possible outcomes. There are three types of nodes in a decision tree:

1. A root node, it has no incoming edges but it has zero or many outgoing edges.
2. Internal nodes, each of the node has only one incoming edge but it has two or more outgoing edges.
3. Leaf nodes, each of the node has only one incoming edge but it does not have outgoing edges. Also, each leaf node has a class label attached to it.

They are different types of nodes in a Decision tree. The main objective of a decision tree classifiers is:

1. It used to classify precisely as much possible training sample.
2. Generalize beyond the training samples so that unseen samples could be classified with high accuracy
3. Easy to update as more training samples become available
4. Have as simple structure as possible

Decision tree implementation [4] are algorithms that can creates a decision tree from a given dataset or training set. There are many number of decision trees developed from a given training set and samples. However, by using those training set and samples, retrieving an optimal tree is an impossible task. A few algorithms have been introduced to get accurate result from decision tree in a reasonable amount of time. These algorithms usually use a greedy strategy to develop a decision tree by creating a series of locally optimum decisions. This consider as attribute that use to partition the data.

The above-mentioned algorithm pointing the following two important questions:

1) How to split the training data?

In each recursive step of the growing tree process, an attribute should be selected for test condition to divide the data into smaller subsets. In order to implement this step, we need to implement algorithm that can provide a method for specifying the test condition for different attributes types, as well as create an objective measure for evaluate each test condition.

2) How do we stop the splitting procedure?

We can stop the splitting condition when the

**4) APPROACH**

The research approach of our project is described in the following steps.

**Step 1: From the given Cashier Risk Report: Grouping Safeway stores in to categories:**

a) Stores having large number of high Risk Cashiers.

b) Stores having low number of high Risk Cashiers.

**Step 2: Identify the location of each store. (Store number is given in the report)**

Location information for Safeway stores is obtained from local.safeway.com. We   
extracted the store address of each Safeway store. And later we mapped the store number   
of a store to the city or area of the store.

**Step 3: Obtaining information about the city or area of location.**

Based on the city lists, we retrieve population of the city, average income and the cashier’s salary (hourly) for each city.

**Step 4: Identify relations between the external factors and the no. of risk cashiers in a store.**

Create a model between the number of high risk Cashiers in a store data from the spreadsheet and the city information that we collect through other sources.

**Step 5: Check the findings on the test data**

Our main objective is to interpret the relation between the risk factor from each store   
and city information such as population, average incomes, cashier’s salary.

**5) RESULTS AND EVALUATIONS**

One of the objectives of the project was to collect data. We collected the store locations for all the stores listed in the Safeway Risk analysis report from Safeway Store Locator [2]. In total, we found 237 unique locations. We couldn’t find locations for all the stores listed in the report. We didn’t use the missing data for this project. A map of all the identified locations is shown in Fig 1.

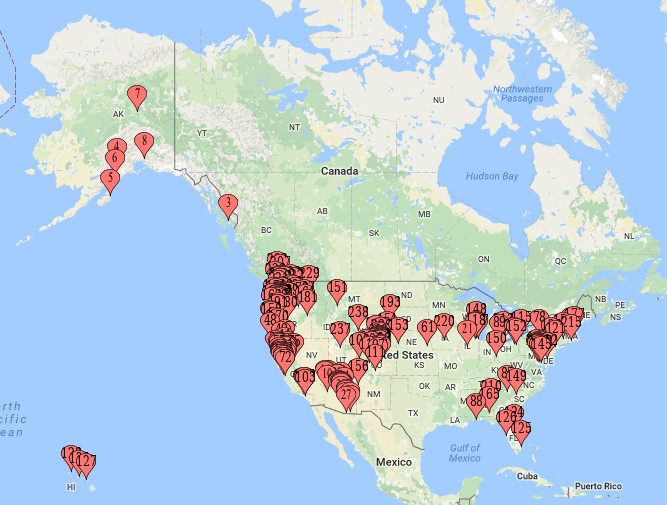
****

Fig 1: Cities having Safeway Store

We obtained the location wise information like population and median household income from [www.city-data.com](http://www.city-data.com) [3]. The complete data was formulated by aggregating all the data collected and identifying the number of risky employees for each store. The data used for the project is shown in Safeway Store Data.xls. The location data collected for the project is given in City Data.xls.. A plot for Store No. vs No. of Risky Cashiers is shown in Fig 2 and a plot of population vs median household income for all the identified stores is shown in Fig 3.







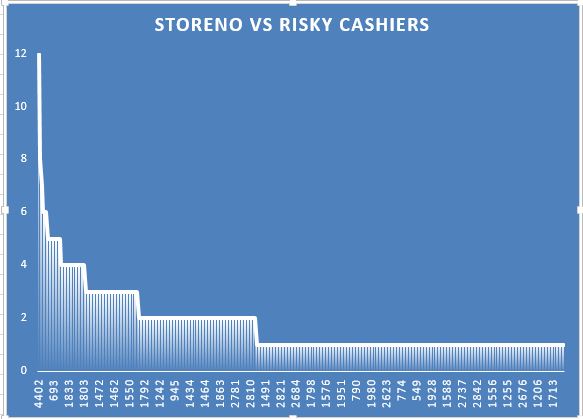


Fig 2: Store No. vs No. of Risky Cashiers

Fig 3: Household Income Vs Population

In this project, we labeled the gathered data in three different ways. Using these 3 different labels we proceeded with the classification process using decision trees.

**Case 1: The Stores Having More Than 1 Risky Cashiers (Class 1) & The Stores Having Exactly 1 Risky Cashiers (Class 2)**

The stores in the collected data were divided into two categories.

1. The stores having more than 1 Risky employee was categorized as Class 1 and
2. The stores with 1 Risky employee was categorized as Class 2.

The data was divided into 2 parts. 60% of the data in each class was used as a training data to train the decision tree. The remaining data in each class was used to test the decisions made by the trained decision tree. A plot of training data is shown in Fig 4. A plot of the test data is shown in Fig 5.

Fig 4: Case 1 Training Data

Fig 5: Case 1 Test Data

A graphical representation of the decision making of the decision tree is placed in Graph 1.pdf file.



In this case, the decision tree could accurately classify 44% of the total test data.

**Case 2:** **The Stores Having More Than 2 Risky Cashiers (Class 1) & The Stores Having Up To 2 Risky Cashiers (Class 2)**

The stores in the collected data were divided into two categories.

1. The stores having more than 2 Risky employees was categorized as Class 1 and
2. The stores having up to 2 risky employees was categorized as Class 2.

Like case 1, the data was divided into 2 parts. 60% of the data in each class was used as a training data to train the decision tree. The remaining data in each class was used to test the decisions made by the trained decision tree. A plot of training data is shown in Fig 6. A plot of the test data is shown in Fig 7.

Fig 6: Case 2 Training Data

Fig 7: Case 2 Test Data

A graphical representation of the decision-making process of the trained decision tree is shown in Graph 2.pdf.



In this case, the classifier could classify 62% of the data correctly.

**Case 3: The Stores Having More Than 2 Risky Cashiers (Class 1) & The Stores Having Up To 2 Risky Cashiers (Class 2)**

The stores in the collected data were divided into three categories.

1. The stores having more than 2 to 6 Risky employees is categorized as Class 1,
2. The stores having up to 2 risky employees is categorized as Class 2 and
3. The stores having more than 6 risky employees is categorized as Class 3.

Like the above two cases, the data is divided into 2 parts. 60% of the data in each class is used as a training data to train the decision tree. The remaining data in each class is used to test the decisions made by the trained decision tree. A plot of training data is shown in Fig 8. A plot of the test data is shown in Fig 9.

Fig 8: Case 3 Training Data

Fig 9: Case 3 Test Data

A graphical representation of the decision-making process of the trained decision tree is shown in Graph 3.pdf.



The results obtained from different cases are shown in Table 1.

|  |  |
| --- | --- |
| Case No. | Accuracy |
| Case 1 | 0.449367 |
| Case 2 | 0.625899 |
| Case 3 | 0.604316 |

Table 1: Results obtained for all the cases

**6) OBSTACLES & WORKAROUNDS**

The given Cashier Risk Report consists of dirty data. We could find 620 records of  
genuine data. Based on these 620 records of data we continued further in accomplishing our tasks. While scrapping data from www.city-data.com. Our IP address was blocked from  
accessing the website, so we reinstated the router so that it would give a new IP address. There were a few cities for which the average household income  
was not available. We tried to retrieve the salary information from the Glassdoor and Indeed websites. These websites have their own API which help to get the information in JSON format, it requires registering and valid token. When we tried to register, the website did not work and  
did not provide us a token. Since there are only limited stores in the Safeway risk  
factor report, we must consider only the cities available in the report, also external  
data has salaries of other positions in the Safeway as well. As we are unable to retrieve the salary information from the Glassdoor API and Indeed API we eliminated this factor in our project.

**7) FUTURE WORK**

For this project, we have a few suggestions that might help to obtain better results:

1. Improved data set: In this project, one of the major limitations was missing data and lack of training data. To make an accurate estimate we can request for accurate data for all the Safeway stores in the country.
2. Pruning of Decision Tree: In this project, we used Scikit learn Decision Tree module. This module didn’t have the pruning capability and therefore, the trained tree overfitted the training data. Hence, it was not a generalized classifier.
3. Adding additional external factors: Currently, we are classifying the stores depending upon two features: population and median household income. In future, we can add more features like: average income of the cashier at that location, Average age of the cashiers in a store, Average training score of the cashiers in a store (If calculated).

**8) REFERENCES:**

[1] Given Safeway Risk Report Data

[2] <http://www.local.safeway.com/> (Safeway Store Loacator)

[3] www.city-data.com, Advameg, Inc

[4] Rokach, Lior, and Oded Maimon. ”Decision trees.” Data Mining and Knowledge Discovery Handbook. Springer US, 2005. 165-192. (For Cashier Salary)

[5] <https://www.analyticsvidhya.com/blog/2015/10/beginner-guide-web-scraping-beautiful-soup-python/>

[6] <https://en.wikipedia.org/wiki/Web_scraping>

[7] https://en.wikipedia.org/wiki/Decision\_tree