PalArch's Journal of Archaeology of Egypt / Egyptology

Data Mining Model to Analyse Crm in Banking Sector

Dr.N.Kannaiya Raja[1], Dr. Kaja Masthan[2], Dr. Balachandra Pattanaik[3], Dr.B.Barani Sundaram[4]

Mr.Balam Suresh Kumar[5], Mr.Elangovan.B [6]

[1] Professor,Computer Science Dept.School of Informatics and Electrical Engineering,Ambo University, IOT Campus,Ambo, Ethiopia

[2] Dept of Information technology,College of Informatics,Bule Hora University.Ethiopia.

[3] Professor/ Associate Dean, Department of Electrical and Computer Engg., Blue Hora University, Blue Hora, Ethiopia

[4]Associate Professor, Department of Computer Science,College of Informatics,Bule Hora University,Bule Hora Ethiopia.

[5]lecturer,Department of Electrical and Computer Engg., Blue Hora University, Blue Hora, Ethiopia.

[6] Department of Computer Science,College of Informatics, Bule Hora University, Bule Hora, Ethiopia.

Email: [1]kannaiyaraju123@gmail.com, [2]kaja.masthan98@gmail.com, [3]balapk1971@gmail.com,[4]bsundar2@gmail.com,[5]sureshnit415@gmail.com,[6]elan77@gmail.com

Bornali Borah, Tulika Hazarika: Data Mining Model to Analyse Crm in Banking Sector -- Palarch's Journal Of Archaeology Of Egypt/Egyptology 17(7). ISSN 1567-214x

Keywords: Customer relationship management (CRM), Data mining, Knowledge Discovery in Databases(KDD), Support vector machine(SVM), recency, frequency, and monetary (RFM), K-means, Logistic Regression(LR)

ABSTRACT

The Banking business is profoundly serious, dynamic and subject to quick change. Thus, Banks are being pushed to comprehend and rapidly react to the individual needs and needs of their clients. Client relationship the executives (CRM) is the general procedure of abusing client related data and utilizing it to improve the income stream from a current client. Information mining methods are utilized to remove significant client data from accessible databases.

The goal of the research is to propose an efficient Customer Relationship Management data mining model for the prediction of customer relationship in the domain of banking applications, specifically to profile loyal and lost bank customers based on their RFM (Recency, Frequency and Monetary) score that help to address the problem domain. The subject of this research is Commercial Bank of Ethiopia customer transaction and demographic information of more than 70,000 customers. Within the built model, different segments of customers are studied and evaluated based on their common attributes since customer grouping is the main part of customer relationship management. Customer classification and grouping is done by combining the Support Vector Machine (SVM) ensemble classifier Random Forest and K-means methods along with the application of RFM (recency, frequency, and monetary) model. The prediction of customers is also done using logistic regression model. The outcomes from this examination were empowering, which fortified the conviction that applying information mining strategies could undoubtedly bolster Customer Relationship Management exercises at Commercial Bank of Ethiopia. Later on, more division contemplates utilizing segment and other client's data and utilizing other grouping and arrangement calculations could yield better outcomes.

1. Introduction

Ethiopian financial area is right now contained a national bank (The National Bank of Ethiopia or NBE), two government possessed banks and sixteen private banks. NBE controls the bank's base store rate, in view of the most as of late information, the CBE assembles in excess of 60 percent of absolute bank stores, bank credits and unfamiliar trade. The state-possessed CBE rules the market as far as resources, stores, bank offices, and complete financial workforce. The other government-possessed bank is the Development Bank of Ethiopia (DBE), which gives credits to speculators working in need divisions. The historical backdrop of the CBE goes back to the foundation of the State Bank of Ethiopia in 1942. CBE was lawfully settled as an offer organization in 1963. From that point forward, it has been assuming significant jobs in the improvement of the nation. Pioneer to acquaint present day saving money with the nation it has in excess of 1340 branches extended the nation over. The main African keep money with resources of 565.5 billion Birr as on June 30th 2018. CBE Plays a reactant job in the monetary advancement and improvement of the nation, likewise the main bank in Ethiopia to present Automated teller machine (ATM) administration for nearby clients. At present CBE has in excess of 20 million record holders and the quantity of Mobile and Internet Banking clients additionally arrived at more than 1,736,768 as of June 30th 2018. Dynamic ATM card holders arrived at more than 5.2 million (Ethiopia nation business control, 2018).

With the early rapid development, banks of our country have grasp a large number of user groups, have accumulated a lot of customer data, how to dig out useful information from these complex customer data ,how to research on different group of bank customers and analyze their difference in consumer behavior, and making different marketing strategies for them. These all have immeasurable meaning for the bank business model to change from extensive model into fine model.

In business today, it is critical to have the option to fulfill client's needs and needs in light of the fact that the current client decides the heading by picking, in an independently directed way. In the event that an association can't fulfill the prerequisites of the clients, the clients will switch item or specialist co-op promptly which makes losing of the chance and rivalry. It additionally impacts association pay in light of the fact that in the current business atmosphere, numerous organizations go after same clients who have more option to pick than the association. So as to the organization can keep a current client base, they have to comprehend the conduct of clients. The organization must characterize an unmistakable client division to build up associations with clients [1].

Data mining has a supporting role to fulfill the BI goal of providing useful information to decision makers. Before information can be presented to end users, the information must be discovered from the data [2].

2. Related Works

Yong Wang and Dong Sheng Wu [3] expounded the composition and major function of the bank's CRM, and constructs decision tree to analyze the kind of the bank's customers by applying the ID3 algorithm. They used decision tree to attain the intellectual need in the CRM interactive process, help the bank understand the behavior of the customers to a fuller extent, and improve the service level of the bank.

Bahari et al. (2015) [4] proposed a productive CRM-information digging structure for the forecast of client conduct. CRM-information mining structure is useful to oversee relationship among associations and clients. The model improves the dynamic procedure for holding esteemed clients. Information mining procedures like characterization are for the most part utilized in CRM. Bahari et al. thought about two arrangement techniques, Naïve Bayes and Neural Networks and results show that Neural Networks execution is better than Naïve Bayes. The creators applied model on bank showcasing dataset that is standard UCI datasets.

C. Sunil Kumar et al (2015) [5]presented the advantages of applying data warehousing and data mining (DWDM) techniques in CRM of the financial divisions like banking.

Surachai and Anongnart Srivihok (2013) [6] used the principles of data mining to cluster customer segments by using K-Means algorithm and data from web log of various Internet banking websites. Consequently, the results showed that there was a clear distinction between the segments in terms of customer behavior.

Griva et al. (2014) [7]analyze an information mining based system to recognize shopping designs. The achievement of any business relies upon the capacity to comprehend its clients. Understanding the reasons purchasers enter their favored stores is assuming a significant job in accomplishing upper hand and holding their pieces of the overall industry. Today, Business Analytics are useful to investigate the gigantic measure of information so as to pick up clients bits of knowledge and improve client connections. The creators propose

the information mining based system which could be utilized to find designs in clients' visits to a general store and recognize their shopping missions. The use of proposed structure is applied in retail location information of eight agent stores of a Greek retailer. It could be utilized to help a few choices in the retail area and improve the connections among retailers and purchasers. This data is useful to help a few choices in the retail area and improve the connections among retailers and shoppers.

D'Haen et al. (2013) [8] foresee the client productivity during procurement and locate the ideal blend of information source and information mining procedure. The client securing process is a difficult errand for retailing operators. There are models that help them in picking right client to follow. Two significant components of this procedure are benefit of client and likelihood of client once the lead is in reality a client. The creators center around the productivity component. Web information and industrially open information are read for prescient execution and estimated the precision which is most noteworthy. Various information digging strategies are applied for prescient execution like choice trees, calculated relapse and sacked choice trees. As indicated by results stowed choice trees are continually higher in exactness and Web information is improved in anticipating gainfulness than business information. Whenever consolidated both at that point result is far and away superior. The creators characterized an edge to gauge benefit. On the off chance that business volume is higher than in a specific worth that was determined by the mail request arranged organization then it is considered as gainful.

Duchessi et al. (2013) [9] break down ski resort's effect on deals and propose special and promoting procedures utilizing choice tree. A ski resort is a hotel produced for snowboarding, winter sports and skiing, Ski resort utilize diverse specialized techniques like advertising, publicizing and deals advancement to speak with their key market portion. Ski resorts are utilizing another computerized correspondence channels for M-business, the expansion of advanced gadgets and E-trade. The creators propose choice tree models to profile the innovations and administrations. The advances and administrations fuse small scale blogging administrations, resort sites and online coupon administrations. Ski resorts sections client in two significant classifications for example millennial or Generation Y (not exactly or equivalent to 35) and non-millennial (more noteworthy than 35) and use advancing and promoting for them. Ski resorts incredibly utilize Social Media Networks, Websites and Micro blogging advances for advancing and promoting their administrations. As per result the innovations and administrations have extraordinary effect on resort deals. The effect is empowering and commonly moment and relentless in nature.

3. Methodology

The main challenge in customer segmentation is the presence of different customer related variables and the high degree of heterogeneity in customer

behavior. Customer variables are generally divided into two categories: behavioral variables and characteristic variables. Characteristic variables included demographic, geographic and psychological variables, while behavioral variables encompassed reaction and tendency of customers to services.

The characteristic variable used in the study is demographic variable i.e. Age and the behavioral variables are Recency, Frequency and Monetary value of the customer.

Based on the above related works, this research wok adopted the clustering task using K-Means algorithm and classification rules using SVM algorithm and classifying the customer segment using RFM model to examine valuable customers. The prediction of customers' behaviour is done using logistic regression model. Model building and analysis was done using python 3.5.1.

3.1. Data preparation

The data preparation phase covers all activities performed to construct the final dataset that could be submitted to the tool from the initial raw data. The tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection as well as transformation and cleaning of data for modeling tools.

3.2 Modeling

a) Segment RFM variables for each customer

In this progression, RFM investigation is applied by characterizing the scaling of R, M, and F attributes. RFM is an acronym for Recency, Frequency, and Monetary. **Recency** is about when was the customer used the service from the bank. It means the number of days a customer made the last service.

Frequency is about how many times a customer used a service from the bank in a given time. A given time I have used in the study is an annual transaction record of the customer.

Monetary is about the total amount of money a customer spent in the bank in a given time.

By applying RFM model to discover client steadfastness, clients are grouped into 5 classes utilizing k-implies calculation and allocate kinds of clients. Mastermind the three characteristics in either rising or plunging request. The score 5 which spoke to "extremely high" is relegated to most contributing one then 4 which spoke to "high" to next most noteworthy contributing one, etc till 1. The score 5 speak to most client dedication and 1 speak to least client reliability. This, there are complete RFM Array 125 (5x5x5) blends since each quality in R, F, M properties has 5 scales (5, 4, 3, 2 and 1).

b) Customer clustering by K-means

The methodology in this stage are that: initially, client devotion is divided by the 5 groups utilizing the K-implies calculation and afterward decide client unwaveringness to each class, which are "High" , "High" , "Medium" , "Low" and "Extremely Low". At that point, the framework will rehash the initial step for every client dependability as per the kind of client.

c) Classification using SVM

The principles of order are found from the three client exchange factors, client devotion (class), and utilizing the grouping results acquired in the past advance. SVM gathering learning method for CRM is proposed to distinguish dependability of client to the bank in subtleties.

Ensemble learning algorithms use an ensemble or a group of classifiers to classify data. Hence they give more accurate results as compared to individual classifiers. RFC is an example for ensemble classifiers. RFC make use of an ensemble of classification trees (Tarun Rao, 2014).

In the study SVM individual classifier and RFC ensemble classifier is used to increase the performance of the classification accuracy.

d) Association rule mining

Association rule mining is a popular data mining method. The data we have used for association consists of the annual transaction of customers based on variables R, F, and M. For each transaction, there is a list of RFM. An association rule is a statement of the form $(\text{item set } A) \Rightarrow (\text{item set } B)$. The point of the examination is to decide the quality of all the affiliation rules among a lot of things. The quality of the affiliation is estimated by the help and certainty of the standard. The help for the standard $A \Rightarrow B$ is the likelihood that the two thing sets happen together. The help of the standard $A \Rightarrow B$ is evaluated by the accompanying:

$$\frac{\textit{transactions that contain every item in } A \textit{ and } B}{\textit{all transactions}}$$

Notice that help is symmetric. That is, the help of the standard $A \Rightarrow B$ is equivalent to the help of the standard $B \Rightarrow A$. The certainty of an affiliation rule $A \Rightarrow B$ is the restrictive likelihood of an exchange containing thing set B given that it contains thing set A. The certainty is assessed by the accompanying:

$$\frac{\textit{transactions that contain every item in } A \textit{ and } B}{\textit{transactions that contain the items in } A}$$

e) Prediction using Logistic Regression

Logistic regression is the most essential analytic tools that use discriminative classifier. In the study logistic regression is used to classify an observation into ‘positive sentiment’ and ‘negative sentiment’. Using logistic regression we use four stages:

1. An element portrayal of the info. For each information perception x(i), this will be a vector of highlights [x1, x2,… , xn]. We will by and large allude to highlight *i* for input $x^{(j)}$ as $x^{(i)}j$, some of the time rearranged as x_i,, yet we will likewise observe the documentation f_i, $f_i(x)$, or, for multiclass characterization, $f_i(c, x)$.

2. A grouping capacity that registers y, the assessed class, by means of p(y/x). In the following area we will present the sigmoid apparatuses for characterization.

3. A target capacities for learning, for the most part including limiting blunder on preparing models. We will present the cross-entropy misfortune work

4. A calculation for upgrading the goal work. We present the stochastic inclination plunge calculation.
Logistic regression has two phases:
Training: Train the framework (explicitly the loads w and b) utilizing stochastic angle drop and the cross-entropy misfortune.
Test: Given a test model x figure p(y/x) and return the higher likelihood name y = 1 or y = 0.

4. EXPERIMENTAL RESULT

This research used database for customers who have an account and active users of CBE, Farisi Branch between 1st June 2018 and 1st June 2019 totaling to 70,000. In the wake of making a choice of information, the records which incorporate missing qualities and erroneous qualities are evacuated, and the excess traits are killed. Next, the information is changed into proper arrangements. At last, the dataset stays 67,997 examples which are described by the accompanying 12 attributes are Sex, age, marital status, year of join, occupation , address , type of account, ATM, POS, Recency amount of each customer transaction, Frequency amount of each customer transaction, Monetary amount of each customer transaction.

4.1 RFM analysis

Customer segmentation is the technique of diving customers into groups based on their transaction patterns to identify who are the most profitable groups. In segmenting customers, various criteria can also be used depending on the market such as geographic, demographic characteristics or behavior bases. This technique assumes that groups with different features require different approaches to marketing and wants to figure out the groups who can boost their profitability the most (Jiwon Jiwong, 2019). The dataset used for research is the transaction history data occurring from 1st June 2018 and 1st June 2019. Based on the transaction dataset RFM analysis is implemented on the dataset. The first thing In the RFM analysis is to count the recency value, the number of days since the last transaction of a customer. From the transaction date column we can get when was the first transaction and when was the last transaction of a customer. As transaction date column also contains transaction time data, the year, month and day part is only extracted. The final day of our dataset is June 1st, 2019. Therefore set July 1st as our pining date and count backward the number of days from the latest purchase for each customer. That will be the recency value.
To count the monetary value each customer spent, the monetary values of all transactions added in each row of the customer. To count the frequency the transaction date column for each customer is used. Based on the count it is simple to count how many times the customer visited the bank on the specified time period. Now the data for each customer is grouped and aggregated for each recency, frequency, and monetary value.

Cid	Recency	Frequency	Monetary
100	211	67	1200
123	226	73	1000
143	240	79	10000
165	255	85	5896
178	269	91	2354
99	284	98	3568
80	298	104	45879
140	47	110	6352
112	30	116	5698
145	19	122	365241
141	60	129	526314

Table 1: Sample RFM value of customers

As shown on the above table customers are grouped based on RFM values and given RFM score depending on the count amount. Customers with low recency amount have high Recency score, customers with high frequency value have high frequency score, and customers with high monetary value have high monetary score.

Cid	Age	R	F	M	RFM score
1	Adult	2	2	1	221
2	Adult	1	1	1	111
3	Youth	2	2	2	222
4	Youth	1	2	1	121
5	Adult	2	2	1	221
6	Youth	2	3	1	231
7	Old	3	2	1	321
8	Youth	3	2	2	322
9	Adult	3	3	3	333
10	youth	2	3	2	232
11	Old	3	3	1	331
12	Adult	3	2	2	322

Table 2: Sample RFM score of customers

4.2. Classification using ensemble SVM

A real-world CBE customer's transaction data set is used to test the performance of the SVM based ensemble data mining model. The data set comprises detailed information of 69,997 customers. The study used five stages for building a SVM-based ensemble learning system. The stages are: data preparation, individual classifier construction, ensemble member selection, ensemble classifier construction, and model evaluation.

```
In [106]: from sklearn.metrics import classification_report, confusion_matrix
          print(confusion_matrix(y_test,y_pred))
          print(classification_report(y_test,y_pred))
```

```
[[3680    0    0    0    0]
 [   0 1208    0    0    0]
 [   0    0 2263    0    0]
 [   0    0    0 5404    0]
 [   0    0    0    0 1445]]
              precision    recall  f1-score   support

        High       1.00      1.00      1.00      3680
         Low       1.00      1.00      1.00      1208
      Medium       1.00      1.00      1.00      2263
   Very High       1.00      1.00      1.00      5404
    Very Low       1.00      1.00      1.00      1445

   micro avg       1.00      1.00      1.00     14000
   macro avg       1.00      1.00      1.00     14000
weighted avg       1.00      1.00      1.00     14000
```

4.3Association rule mining

A rule consists of a left-hand side proposition (antecedent) and a right-hand side (consequent).

An affiliation rule is a standard as X →Y, Where X and Y are predicates or set of things. As the quantity of created affiliations may be enormous, and not all the found affiliations are important, two likelihood measures, called backing and certainty, are acquainted with dispose of the less incessant relationship in the database. The help is the joint likelihood to discover X and Y in a similar gathering; the certainty is the restrictive likelihood to discover in a gathering Y having discovered X (). Association rules in the study are generated by using the class (loyalty) generated from the RFM score analysis. The rules contain (Medium, High) (High, Medium) and (Low, High) antecedents and consequents. The support value for the first rule is 0.510765. This number is calculated by dividing the number of transactions containing medium loyalty divided by total number of transactions. The confidence level for the rule is 0.890283 which shows that out of all the transactions that contain medium loyalty, 89.02% of the transactions also contain high loyalty.

	antecedents	consequents	antecedent support	consequent support	support
0	(Medium)	(High)	0.573710	0.834021	0.510765
1	(High)	(Medium)	0.834021	0.573710	0.510765
2	(Low)	(High)	0.492950	0.834021	0.338386

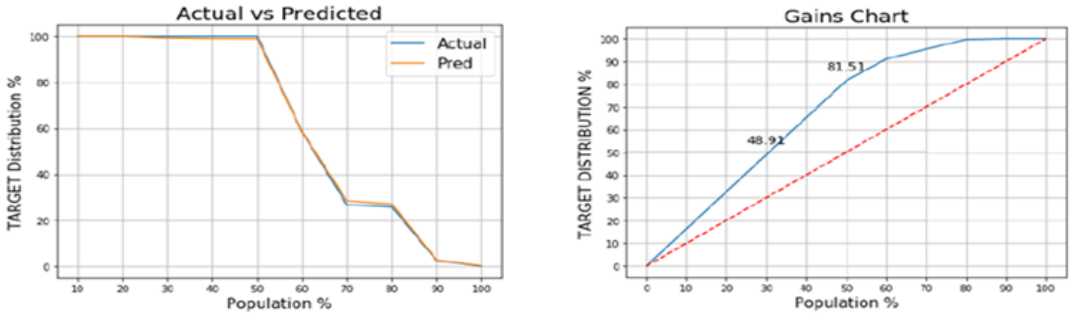
	confidence	lift	leverage	conviction
0	0.890283	1.067459	0.032278	1.512793
1	0.612412	1.067459	0.032278	1.099853
2	0.686451	0.823062	-0.072745	0.529355

4.4. Logistic regression for prediction

The task is to find a functional form for *f* that can correctly predict new cases that the SVM has not been presented with before. This can be achieved by training the SVM model on a sample set, i.e., training set, a process that involves, classification, the sequential optimization of an error function. In this logistic regression model, a step by step scikit learn library in Python is used.

The first step is loading the dataset using Pandas library. Increase and Lift outlines are utilized to assess execution of forecast model. They measure how much better one can hope to do with the prescient model contrasting without a model. It's an extremely well known measurement in showcasing investigation. It's not simply limited to advertising investigation. It very well may be utilized in different areas also, for example, chance displaying, gracefully chain examination and so forth. It likewise assists with finding the best prescient model among various challenger models.

A) Actual Vs Predicted result



5. Conclusion

This research attempted to study the possible application of data mining techniques, and especially RFM and SVM with clustering techniques, to support CRM at CBE. The study was conducted in three major phases, namely business understanding, model building, and evaluation.

The data collection and preparation were major tasks due to the dispersed nature of the required data. Business understanding includes the CRM parameters of customer identification, customer attraction, customer retention and customer development. In model building phase data preparation and preprocessing have been done that helped in the identification of RFM values for each customer, which made the experiment lengthy. Next, K-means clustering algorithms were applied to segment the customer data into meaningful groups. The parameters used by K-means algorithm were defined, based on the RFM score. The use of RFM analysis helps improving customer relationship even better than by directly going for data mining since it incorporates customer demographic variables as well in getting the results. Thus in the competing world of today RFM analysis helps organizations to better attain their goals of profit and customer relationship.

Although it is difficult to generalize based on these results, findings of the study seem to validate the business norm that ‘customer value’ is based on the total monetary value spent in the Bank. The cluster model, which according to the domain experts made business sense, segmented the records into five clusters. Two of the clusters, 5&4 contained 74.2% of customer records that generated the highest RFM score of customers. The cluster containing medium and low RFM score generating customers contained 7.4% and 18.42% of the customer records used in the study respectively. SVM linear kernel is a relatively efficient classification method used to classify bank's customers.

Based on the training, an ensemble classifier RF result accuracy of 92% that shows an ensemble RFC has best performance for classifying. After selecting the train and test datasets, the Prediction from the logistic regression resulted accuracy 92.8% that is a very good performance. It becomes easier to implement bank's customer relationship strategies with SVM-based binary classification model, which allows bank to attract more high-quality customer resources and reduce the loss of existing customers. It also has a significant reference value for the design of the bank's customer relationship management system. In addition to confirming current business knowledge, the clusters provided a new view of customer segments with different transaction behavior. In general, the results from this study were encouraging. It was possible to segment customer data using data mining techniques that made business sense. Associations should better comprehend their clients. Particularly, it is fundamental for organizations, they ought to have point by point understanding about their client's characteristics, behaviors, demographics, etc. It is the researcher's belief that a more thorough study using data mining techniques can increase business leverage from customers and support CRM activities at Ethiopia. Furthermore, knowledge of data mining techniques, marketing strategies and Banking business processes should be integrated to successfully implement CRM.

Reference

A. Griva, C. Bardaki, S. Panagiotis, and D. Papakiriakopoulos, "A Data Mining Based Framework to Identify Shopping Missions," 2014.

Bharati M. Ramageri, "DATA MINING TECHNIQUES AND APPLICATIONS", Indian Journal of Computer Science and Engineering Vol. 1 No. 4

Bobinski, G.S. and Assar, A. (1994), "Division of financial responsibility in baby boomer couples: routine tasks versus Investment:, In Costa, J.A. (Ed.)., Gender Issues and Consumer Behavior, Sage Publications, London.

Carbone P L 2000 "Expanding the meaning of and applications for data mining". In IEEE Int. Conf. On Systems, Man, and Cybernetics 1872–1873.

Carbone P L 2000 "Expanding the meaning of and applications for data mining". In IEEE Int. Conf. On Systems, Man, and Cybernetics 1872–1873.

C. Sunil Kumar ,P.N. Santosh Kumar ,T. Venkata Mohit ,A. Mahesh, Data Mining Techniques for Banking Applications, International Journal of Research Studies in Computer Science and Engineering (IJRSCSE) Volume 2, Issue 4, April 2015, PP 15-20

Coumaros, J.; Buvat, J.; Auliard, O.; Roys, S.; Kvj, S.; Chretien, L.; Clerk, V. Big Data Alchemy: How can banks maximize the value of their customer data. In Banks Have Not Fully Exploited the Potential of Customer Data; Digital Transformation Research Institute and Capgemini Consulting: Paris, France, 2014.

Dimitris Christodoulakis (2017) RFM analysis for decision support in e-banking area, Computer Engineering and Informatics Department University of Patras GREECE.

Gattari, P. (2012). Role of Patent Law in Incentivizing Green Technology, The.Nw. J. Tech. & Intell. Prop., 11, vii.