Predicting Purchased Policy for Customers in Allstate Purchase Prediction Challenge on Kaggle

Saba Arslan Shah and Mehreen Saeed
FAST-NU
Lahore, Pakistan.
sabarslan@gmail.com

FAST-NU
Lahore, Pakistan.
mehreen.saeed@nu.edu.pk

Abstract

This paper gives an overview of the methodology developed for predicting the purchased policy for a customer in Allstate purchase prediction challenge held by Kaggle (Kaggle). It gives an account of challenges faced during the process and strategies used to predict the policy choices for customers. The techniques used include logistic regression, naïve Bayes (Mitchell, 1997), SVM (Lin., 2011) , random forest (Breiman, 2001), probability calculation for each policy and its change and a voting mechanism. Effect of previously presented policies is also measured. The dataset presented a challenge since it included feature set in both rows and columns for each of the customer. Furthermore, seven policy options were to be correct as a combination, for a prediction to be deemed accurate. Relationships are also explored between different policy options.

Keywords: *Purchased policy prediction, Allstate purchase prediction challenge, SVM, Data mining, Random forest*

1. Introduction

Allstate purchase prediction challenge was hosted by Kaggle from February, 2014 to May, 2014. It consisted of 1571 participating teams from all over the world. The challenge required to predict the car insurance policy a customer buys after receiving a number of quotes. The insurance policy is a group of seven individual policies each with two to four values. For a prediction to be correct, all seven options have to be correctly predicted for each customer. This represents a classification problem with more than 2000 expected classes' altogether. The data consists of both training and test set. Training set consists of customer data along with the purchased policy information, while test data consists of different customers, and does not have a value of purchased policy since it has to be predicted. Test set data is also truncated to make the competition more challenging. Training and test set data had the same set of features present. For calculation of the accuracy score of the predicted policies, 30% of the test set was used for public leaderboard.

Predicting consumer's purchase behavior has been studied extensively by researchers, of which, recommender systems is one example (Schafer, 2001). Random forests and regression forests have been used to predict next buys for customers who are shopping for financial products. It is argued that past customer behavior impacts the next bought financial products (Bart Larivie`re, 2005). A kernel-based semi-supervised learning with Laplacian kernel matrices approach is also used with both customer demographic data and their purchase history which was taken from a fairly large dataset (Yajima, 2007). In their study, (Chong Wang, 2012) give an account of a customer's purchase sequence evolution given a period of time. Frequent patter mining is consistently been used in data mining applications (Jiawei Han, 2007). Frequent pattern mining combined with customer's purchase sequence gives better results in predicting customer purchases (Md. Rezaul Karim, 2012). A related approach is used in our research, but our dataset lacked the historical data of a customer's previously

purchased policies, therefore, these approaches have to be modified to suit the given dataset. Our dataset contained shopping points, which represented the policies presented to the customers before they made a purchase decision. We have based our approach on the last viewed policy for a customer, and built probabilistic models and voting mechanism on top of this approach.

The rest of this paper is organized as follows. Section 2 presents the research questions which were formulated at the start of the research. Section 3 and 4 gives an account of exploratory analysis and subsequent data transformations. Section 5 gives a brief explanation of the challenges faced, followed by section 6 which gives a detailed description of different strategies that were used for prediction. Section 7 discusses the strategies and their results and we conclude at section 8.

2. Research Questions

The following research questions were chosen for the purpose of this paper.

- Does customer demographics have any effect on selection of purchased policy?
- Does number of quotes received by a customer plays a role in selection of purchased policy?
- Does shopping policies have any effect on the outcome of purchased policy?

3. Exploratory Data Analysis

All participants were given two datasets. One represented the training set and the other one was test set. Training set contained 665249 rows, each containing multiple policies seen by a customer at different times along with the option they ultimately bought. There were 97009 unique customers and 1809 unique policy combinations present in the training set. Test set contained 198856 rows, each containing policies seen by the customer, but there was no information on which policy did the customer bought in the end. The test set data was truncated and it contained only a partial shopping history of the quoted policy for a customer (Allstate Purchase Prediction Challenge). Test data had 55717 unique customers.

The dataset presented by Kaggle (Kaggle) had 25 features in total, representing customer demographics and other details. These features were then divided into continuous and categorical variables. Some of the categorical variables were transformed into binary data to improve the efficiency of machine learning algorithms running on them. This was achieved through writing data manipulation routines in Matlab.

3.1. Missing Values

Some of the variables had missing values in them; therefore, a substitution was required. The following variables had NA values, indicative of a missing value, and the strategies used to assign values are as follows.

location: Location represented where the shopping point occurred (Allstate Purchase Prediction Challenge). Its missing values were replaced with 0.

risk_factor: This variable assessed how risky a customer is, and it could contain ordinal values from 1 to 4 (Allstate Purchase Prediction Challenge). A prediction model was created, which matched the customer data with other customers. Similar customer's risk factor value was taken into account and the most likely value was given to the customer with NA value. This strategy was used for both training and test set data.

C_previous: This variable represented what the customer had previously bought for product option C (Allstate Purchase Prediction Challenge). The NA values for this option were replaced with a 0, which represented that the customer did not have this product previously.

duration_previous: The variable expressed the duration a customer was covered by their previous issuer (Allstate Purchase Prediction Challenge). The missing values for this option were also replaced by 0, indicating that the customer did not have a policy previously.

3.2. Relationships between Different Plans

Relationships between different plans were observed extensively. All plan options were compared against each other pairwise. Some of the plan options showed positive relationship between different possibilities of values. The relationship graph between different selections of option A and subsequent occurrence of option B for that particular option A value is given in figure 1. Several figures, similar to figure 1, were made to get an overview of the probabilities of joint occurrence of different options.

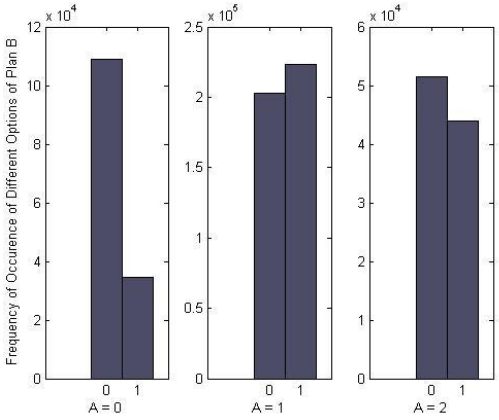

Figure 1: Relationship of Occurrence of Plan B with Different Options of Plan A

4. Data Transformation

The data presented in this competition consisted of both continuous and categorical variables. Table 1 shows the variable names along with their descriptions. Some of the categorical variables were transformed into binary variables with the help of scripts, for both training and test data. The last column shows whether the variable was transformed into a binary variable or not. Variables A to G are treated as response variables whose values were to be predicted.

4.1. Customer bought Policy Data

Since the options A to G represented the response variables and each customer had multiple rows of data representing a bunch of shopping points, with only one row of purchase point in the training set, a dataset was made which contained only the purchased option of each customer. The dataset was named **policy_bought_data**. This dataset was later used in logistic regression to predict individual options.

4.2. Response Variables Value Adjustment

The values for response variables needed to be adjusted to be used with SVM. The multiclass labels were given as starting from 1 and so on. If some values were not there in the training set, they were further adjusted to represent a continuous label value (Lin., 2011).

5. Challenges with Data

Firstly, we did not have data from other companies insurance policies which the customer got quotes on. This could play a part in a customer's mindset of purchasing a particular policy.

Secondly, all seven options had to be predicted correctly for each customer to be counted as a correct entry. Predicting individually correct options had no significance.

Thirdly, the data was not thoroughly explained. The seven coverage options and their meanings were not given. Similarly, car value was given as alphabets without specifying if the alphabets meant something significant. A better explanation might have helped improve prediction.
Lastly, the data was challenging since it had two contrasting features. The columns represented the feature set of a customer, and each customer had multiple rows. It was vital to take care of both of these constraints. Different models were fed with a different features and sets of the data.

Table 1: Variable names, descriptions (Allstate Purchase Prediction Challenge) (http://www.kaggle.com/c/allstate-purchase-prediction-challenge/data) and transformation information

Variable Name	**Description**	**Type**	**Changed to Binary**
customer_ID	Customer's identifier	identifier	No
shopping_pt	Plans presented to a customer - Unique identifier	categorical	No
record_type	Record type can either be 0 = shopping point or 1 = purchase point	categorical	No
Day	Day of the week (0 - 6, 0 = Monday)	categorical	Yes
Time	Time of day (HH:MM)		
State	State where shopping point occurred	categorical	No
Location	Location ID where shopping point occurred	categorical	No
group_size	Number of people covered under the given policy. Ranges between 0 - 4	continuous	No
Homeowner	Whether the customer owns a home or not (0 = no, 1 = yes)	categorical	No
car_age	How old is the customer's car	continuous	
car_value	What was the value of the customer's car when new	categorical	Yes
risk_factor	How risky is the customer. Ranges between 1 - 4	categorical	Yes
age_oldest	Age of the oldest person in customer's group	continuous	
age_youngest	Age of the youngest person in customer's group	continuous	
married_couple	Is anyone in the customer group is married. (0 = no, 1 = yes)	categorical	No
C_previous	What the customer formerly had or currently has for product option C (0 = nothing, 1, 2, 3,4)	categorical	Yes
duration_previous	For how long (in years) the customer was covered by their previous issuer	continuous	
A,B,C,D,E,F,G	Different coverage options	response	Yes
cost	Cost associated with the quoted coverage options	continuous	

6. Model Construction

Matlab (MATLAB, 2014) was used for the purpose of this competition as the tool of choice. Exploratory data analysis was done in R (R Core Team, 2014) and Python's (Rossum, 1995) scikit learn library (Pedregosa, 2011) was used for random forests (Breiman, 2001) implementation. The following strategies were used in the process of building a prediction model. A summary of different strategies and their corresponding accuracy scores is given in table 2 along with the final standings on private leaderboard when 100% of the test data was used to calculate prediction accuracy.

6.1. Strategy 1

First of all, categories A to G were taken into account separately. A prediction model was needed to predict the values of these categories. Options B and E were two class problems. Logistic regression algorithm (Mitchell, 1997) was implemented on these options. The model was fed with all the training data which consisted of the purchased policies of a customer. The dataset policy_bought_data was used for this purpose. This dataset ignored the quoted policies completely. The idea was to see if any correlation existed between the customer data and his purchase habits. After predicting options B and E, the logistic regression model was then extended to be used for multiclass problems, since the rest of the options (A, C, D, F, G) consisted of more than two classes.

A validation set was created using random 30% of the training set. The weight values were adjusted after minimizing the error on validation set. The weights were then applied to the test set and a prediction file was created for each customer in the test set. Submission of this file to the competition leaderboard gave an accuracy of 0.00 on 30% of the test set.

6.2. Strategy 2

On the second attempt, Naïve Bayes algorithm (Mitchell, 1997) was used keeping in view all categorical variables of the customer, ignoring continuous variables. All categorical variables were converted to binary variables as explained in the data transformation potion. Naïve Bayes was applied to all seven categories and prediction score of 0.00 was attained.

6.3. Strategy 3

Third attempt was taken with SVM. Response variable values were adjusted as explained in section 4: data transformation. LibSVM (Lin., 2011) package was used for predicting both two class and multi class labels. The model was adjusted using validation set accuracy. On validation set, the accuracy reached 100% but when the same was applied to test set, the accuracy score remain 0.00 on the public leaderboard. Perhaps, the model was over fitting to the training data.

6.4. Strategy 4

In the next attempt, random forest (Breiman, 2001) was used. This algorithm was applied to the features of a customer and an outcome was predicted for all the seven options (A - G) independently. The number of trees in the forest (n_estimators) in scikit learn (Pedregosa, 2011) was given to be 100 for each of the models. The leaderboard score of 0.02779 was achieved for this method.

6.5. Strategy 5

Fifth attempt made use of random forest (Breiman, 2001) again. This time the number of trees in the forest (n_estimators) in scikit learn (Pedregosa, 2011) was given to be 200 for option A and option G. This model gave a leaderboard score of 0.02856 which was slightly higher than the score of strategy 4.

6.6. Strategy 6 and Last Quoted Policy Effect

Constantly getting a low accuracy score on the test set resulted in altering the strategy. This result suggested that some valuable information is not being taken into consideration in the training of the model. Taking a closer look at the data advocated the fact that the last quoted policy had an impact on the purchased policy. Most of the purchased policies are a result of the last quoted policy. An attempt

was made to predict the last quoted policy as it is. This immediately spiked the accuracy on test set to 0.53793, signifying the fact that this result could be used as a baseline strategy for prediction. (This was later confirmed with the fact that last quoted policy was a benchmark in the competition).

6.7. Introduction of a two way approach

This development led to a two way approach of prediction. Firstly, a prediction was needed as to which policies would change for a given customer. Moreover, another prediction was required which would suggest what would these policies ultimately change to.

6.8. Strategy 7

Keeping in mind recent successful approach of predicting last quoted policy, a new strategy was built in an attempt to better the results. Probabilities of all policies were calculated for each customer. The policy combination with the highest probability was then predicted for that particular customer. This was applied to test set and an accuracy of 0.00138 was achieved.

6.9. Strategy 8

The next strategy was to predict the policies quoted as second last options. The idea was to see if going further up would result in having a better estimate. The result came out to be an accuracy score of 0.40242. It was not possible to move further up than the second last policy, since most of the customers in test set had a maximum of two quoted policy records only.

Table 2: Summary of accuracy achieved on test set using different strategies.

Strategy	**Description**	**Leaderboard Accuracy Score**
Strategy 1	Logistic Regression (Mitchell, 1997) applied on customer characteristics	0.00
Strategy 2	Naïve Bayes (Mitchell, 1997) applied on customer characteristics	0.00
Strategy 3	SVM (Lin., 2011) applied on customer characteristics	0.00
Strategy 4	Random forest (Breiman, 2001) applied to customer characteristics with number of trees = 100	0.02779
Strategy 5	Random forest (Breiman, 2001) applied to customer characteristics with number of trees = 200 for option A and G	0.02856
Strategy 6	Last quoted policy	0.53793
Strategy 7	Probability of policies calculation	0.00138
Strategy 8	Second last quoted policy	0.40242
Strategy 9	Policy's frequency of change	0.53715
Strategy 10	Voting mechanism taking into consideration strategy 1, 3, 4 and 6	0.53733
Final Leaderboard Score	Score of strategy 4 applied on complete test set	0.53266
Winning Strategy	Final score of winner of the competition	0.53743

6.10. Strategy 9

In the next attempt, a given policy's frequency of change is calculated for each customer. It is important to note that the policy was taken as a whole, which means that a combination from A to G

is taken as one complete option for a policy. Individual change to policies was not taken into consideration. There were 1809 unique policy combinations; therefore, an 1809 * 1809 matrix was created, within which each entry represented the number of times a policy as a whole changed to another policy. A count was kept on the number of times a policy changed, and to what policy did it change to. After getting the calculated frequencies, the policy with most change was replaced by the one to which it changed to most frequently. This strategy was applied to only the most frequently changed policy within the last quoted policy prediction set. This yielded a score of 0.53715.

6.11. Strategy 10

In the final attempt, a voting mechanism was created to predict the change of policy from the baseline (last quoted policy). Four predicted results were already calculated. First was the result of logistic regression prediction, second prediction was resultant of SVM. For these two predictions, customer data other than the quoted policy data was taken into account for prediction. The other two predictions were the last quoted policy prediction and second last quoted policy prediction which only considered the different quoted policies presented to each customer. It was assumed that a fusion of both results could produce a better outcome. For each individual item (A to G) a comparison and voting mechanism was developed. The prediction with the highest number of votes was given preference over all others. This strategy resulted in an accuracy score of 0.53733.

7. Discussion and Future Directions

The three research questions posed at the beginning of the problem were partially answered. Customer demographics seem to have little effect on selection of purchased policy. However, more time and effort in this direction could yield fruitful results since the margin of improvement is very thin and slight correlations impacting the policy purchase decision could prove to be fruitful.

The number of quotes received by a customer does play a role in selection of purchased policy. It is seen that with higher number of quotes, the chances of selecting the last quoted policy increases.

The last question on whether the shopping policies have any effect on the outcome of purchased policy is already answered positively. The last quoted policy has a strong impact on the outcome of the results. More than half of the time, it predicts the purchased policy correctly.

This was a tough competition. The margin of improvement was very low. After the competition got finished, the winner had an accuracy score of 0.53743 (as seen in table 2 as “Winning Strategy”) as compared to our best score of 0.53266 (at the rank of 1022), which shows a difference of mere 0.00477 on the private leaderboard. The private leaderboard reflects the accuracy score achieved on 100% test data as opposed to 30% test data of public leaderboard used throughout the competition. The scores of public and private leaderboard are reflected in table 2. The comparison of the winning strategy score and our best score shows that even a slight improvement (maybe a few more correct predictions) in customer prediction contributes to a large jump in the accuracy score and subsequently the rankings. The baseline was last quoted prediction score for a customer. The trick was to see which customers would change to a different policy. In the final solutions, the winning strategy just changed policy G for two states and got an improved overall score.

The test set was truncated which further complicated prediction. Moreover, the data was not explained in great detail e.g. the attribute car value had the categorical data (c, d, e, f, g, h) which did not explain what it represented. A better understanding of the features might have helped in a better prediction.

Correlations with other customer characteristic data can still be investigated to come up with even better solutions. The last quoted policy has a strong impact on the prediction of the policies, therefore, it would be advisable to treat it as a baseline policy and build further models on top of this knowledge.

Frequent pattern mining approach (Jiawei Han, 2007) is another candidate that can be used with shopped and purchased policies to discover relationships between different policy options.

8. Conclusion

This paper gave an overview of the methodology developed for predicting the purchased policy for a customer in Allstate purchase prediction challenge held by Kaggle (Kaggle). It gave an account of challenges faced during the process and strategies used to predict customer's policy purchase choices. More than half of the times, the policy shown to the customer immediately before the purchase decision turned out to be the final purchased policy. Frequency of a policy change and voting mechanism were built on top of this to gauge the effect of this modification to the results.

References

Allstate Purchase Prediction Challenge. (n.d.). Retrieved May 20, 2014, from Kaggle: http://www.kaggle.com/c/allstate-purchase-prediction-challenge/data

Bart Larivie`re, D. V. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications 29* (pp. 472–484). Elsevier Ltd.

Breiman, L. (2001). Random Forests. *Machine Learning*, 5 - 32.

Chong Wang, Y. W. (2012). Discovering Consumer's Behavior Changes Based on Purchase Sequences. *9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2012)* (pp. 642 - 645). IEEE.

Jiawei Han, H. C. (2007). Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, pp 55-86.

Kaggle. (n.d.). Retrieved May 20, 2014, from Kaggle: http://www.kaggle.com/

Lin., C.-C. C.-J. (2011). LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, (pp. 2:27:1--27:27).

MATLAB. (2014). *MATLAB and Statistics Toolbox, Release 2012b[Software]* . Natick, Massachusetts, United States: The MathWorks, Inc.©.

Md. Rezaul Karim, J.-H. J.-S.-J. (2012). Mining E-Shopper's Purchase Rules by Using Maximal Frequent Patterns: An E-Commerce Perspective. *Information Science and Applications (ICISA), 2012 International Conference on* (pp. 1 - 6). Suwon: IEEE.

Mitchell, T. M. (1997). *Machine Learning.* McGraw-Hill Science/Engineering/Math.

Pedregosa, F. a. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2825-2830.

R Core Team. (2014). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rossum, G. (1995). *Python reference manual.* Amsterdam, The Netherlands, The Netherlands ©: CWI (Centre for Mathematics and Computer Science) .

Schafer, J. B. (2001). E-commerce recommendation application. *Journal of Data Mining and Knowledge Discovery*, 16:125–153.

Yajima, Y. (2007). Predicting purchase preferences using semi-supervised one-class SVM with graph kernels. *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on* (pp. 3505 - 3511). Montreal, Que.: IEEE.