Porto Seguro's Safe Driver Prediction
Project Status Report

Hanlin He, Mingze Xu, Su Yang, Tao Wang
Department of Computer Science, The Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas
Email: {hxh160630, mxx160530, sxy161730, txw162630}@utdallas.edu

I. Introduction

Machine learning is emerging in the insurance industry and is being applied across multiple areas including the interpretation of data, business operations and driver safety. One key application is claim prediction. Inaccuracies in car insurance company's claim predictions raise the cost of insurance for good drivers and reduce the price for bad ones. A more accurate prediction will allow insurers to further tailor their prices, and hopefully make auto insurance coverage more accessible to more drivers.

In this report, we based on Kaggle's Featured Prediction Competition *Porto Seguro's Safe Driver Prediction* [1], conducted several experiments, compared different approaches' effectiveness to tackle the claim prediction problem, including *logistic regression*, *tensorflow estimator* and *random forest*.

The organization of the following report is as follows. In section II, we will formally define the problem to solve and discuss the theoretical principle of the algorithm we used. Then we will analyze the data feature and our method of feature engineering in section III. After that, the experimental results are shown and analyzed in section IV. Finally, we will discuss related works and conclude the report.

II. Problem Definition and Algorithm

A. Task Definition

A machine learning problem is defined as to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E [3]. The claim prediction problem can be defined as follow:

- E previous year's policy holders' information and whether or not a claim was filed for that policy holder.
- T predicting the possibility that an auto insurance policy holder will file an insurance claim next year.
- P the accuracies and Gini Coefficient was used to measure the effectiveness of models.

B. Algorithm Definition

Describe in reasonable detail the algorithm you are using to address this problem. A psuedocode description of the algorithm you are using is frequently useful. Trace through a concrete example, showing how your algorithm processes this example. The example should be complex enough to illustrate all of the important aspects of the problem but simple enough to be easily understood. If possible, an intuitively meaningful example is better than one with meaningless symbols.

1) Logistic Regression: Logistic Regression is an approach to learning functions of the form $f : X \rightarrow Y$ [2] or in our case $P(Y|X)$ where Y is discrete-valued, and $X = \langle X_1 ... X_n \rangle$ is any vector containing discrete and continuous variables. The parametric model assumed by Logistic Regression in the case where Y is boolean is:

$$P(Y = 1|X) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i X_i)}$$

$$P(Y = 0|X) = \frac{\exp(w_0 + \sum_{i=1}^{n} w_i X_i)}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i X_i)}$$

One reasonable approach to training Logistic Regression is to choose parameter values that maximize the conditional data likelihood. We also used regularization to reduce the overfitting problem. The penalized log likelihood function is as followed:

$$W \leftarrow \arg\max_{W} \sum_{l} \ln P(Y^l|X^l, W) - \frac{\lambda}{2}||W||^2$$

where the last term is a penalty proportional to the squared magnitude of W.

In general, the algorithm used gradient ascent to repeatedly update the weights in the direction of the gradient, on each iteration changing every weight w_i, beginning with initial weights of zero, according to:

$$w_i \leftarrow w_i + \eta \sum_{l} X_i^l (Y^l - \hat{P}(Y^l = 1|X^l, W)) - \eta \lambda w_i$$

where η is a small constant which determines the step size. The actual implementation of scikit learn library includes multiple solvers, such as Stochastic Average Gradient (SAG) descent, SAGA and Broyden-Fletcher-Goldfarb-Shanno (LBFGS).

III. Feature Analysis and Engineering

The data comes in the traditional Kaggle form of one training and test file each: `train.csv` and `test.csv`. Each row corresponds to a specific policy holder and the columns describe their features. The target variable is named `target` here and it indicates whether this policy holder made an insurance claim in the past.

Feature Engineering.

IV. Experimental Evaluation

A. Methodology

What are criteria you are using to evaluate your method? What specific hypotheses does your experiment test? Describe the experimental methodology that you used. What are the dependent and independent variables? What is the training/test data that was used, and why is it realistic or interesting? Exactly what performance data did you collect and how are you presenting and analyzing it? Comparisons to competing methods that address the same problem are particularly useful.

B. Results

Present the quantitative results of your experiments. Graphical data presentation such as graphs and histograms are frequently better than tables. What are the basic differences revealed in the data. Are they statistically significant?

C. Discussion

Is your hypothesis supported? What conclusions do the results support about the strengths and weaknesses of your method compared to other methods? How can the results be explained in terms of the underlying properties of the algorithm and/or the data.

V. Related Work

Answer the following questions for each piece of related work that addresses the same or a similar problem. What is their problem and method? How is your problem and method different? Why is your problem and method better?

VI. Future Work

What are the major shortcomings of your current method? For each shortcoming, propose additions or enhancements that would help overcome it.

VII. Conclusion

Briefly summarize the important results and conclusions presented in the paper. What are the most important points illustrated by your work? How will your results improve future research and applications in the area?

VIII. Bilbiography

Be sure to include a standard, well-formated, comprehensive bibliography with citations from the text referring to previously published papers in the scientific literature that you utilized or are related to your work.

References

[1] Kaggle Inc. Porto Seguros Safe Driver Prediction, 2017. [Online; accessed 01-Nov-2017].

[2] Thomas M. Mitchell. *Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression*. draft of February, 2016.

[3] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.