Implementation study: Using decision tree induction to discover profitable locations to sell pet insurance for a startup company

Group 9 Members
Department of Computer Science, University of Ghana Legon

Introduction

Context: The paper focuses on an implementation study that utilizes decision tree induction as a methodology to identify profitable locations for selling pet insurance. The study aims to assist a startup company in determining optimal locations by leveraging decision tree analysis techniques.

Objective:The objective of the paper is to investigate the use of decision tree induction as a methodology to identify profitable locations for selling pet insurance by a startup company.

Methodology: The authors used a decision tree induction algorithm to identify profitable locations for a pet insurance startup. They used publicly available data including US Census data and veterinary surgery location data as their data sources. They evaluated their results using the potential profits generated by each of the algorithms as key performance metrics.

Results

1. The authors found that the decision tree induction algorithm was able to identify profitable locations for a pet insurance startup.
2. The algorithm was able to generate a list of the top 10 most profitable locations for the startup.
3. The top 10 locations were all in the United States, and they were all located in areas with high pet ownership rates.
4. The authors concluded that the decision tree induction algorithm was a valuable tool for identifying profitable locations for pet insurance startups.

Conclusion: The decision tree induction algorithm was able to identify profitable locations for a pet insurance startup, with the potential to generate significant profits.

Research Contribution

1. Demonstration of the use of decision tree induction
2. Use of publicly available data
3. Evaluation of results using potential profits.

Problem Definition

•The paper discusses the problem of discovering profitable locations to sell pet insurance for a startup company. The authors mention that businesses generate massive amounts of data, and while standard online analytical processing (OLAP) tools are excellent at performing their reporting function, they are not capable of generating the kinds of insights that businesses require. This has led to substantial research into data mining, with a limited amount of work focused on utilizing large business databases for marketing-related efforts. The authors aim to determine the characteristics of individuals who are the most promising candidates for pet insurance, so that marketing campaigns can be targeted at individuals of this type.

Methods Used

•The paper uses decision tree induction as the method to discover profitable locations to sell pet insurance for a startup company.

Decision trees are effective tools that help to choose between several courses of action. They describe a tree structure wherein leaves represent classifications and branches represent conjunctions of features that lead to those classifications.

A decision tree can be learned by splitting the source data set into subsets based on an attribute value test. The authors use this method to determine the characteristics of individuals who are the most promising candidates for pet insurance, so that marketing campaigns can be targeted at individuals of this type.

SBP Algorithm Method

•The first algorithm used was the SBP algorithm, which is a profit-based algorithm that classifies data with the aim of earning more profit rather than to be more accurate

C4.5 Algorithm Method

•The second algorithm used was the C4.5 algorithm, which is based on the ID3 algorithm and contains several improvements like choosing an appropriate attribute selection measure, handling training data with missing attribute values, handling attributes with differing costs, and handling continuous attributes

Results

- *Bellwether effect* exist in a sample space of *N>100* projects to obtain a potential *Bellwether moving window.*
- Five (5) partition samples with initial approximate size of 219 projects (20%) for ISBSG dataset and 3 partition samples with initial size of 47 projects (33.1%) for Kitchenham dataset.
- There exist an ergodic Markov chain from the TPM showing that the partition samples are stationary.
- With regards to the use of weighting functions, we realized that the Gaussian function yielded superior prediction accuracy.
- In using the *Bellwether moving window* to make prediction on the *new* projects (hold-out), we realized that Deep Neural Networks (DNN) yielded superior accuracy in both datasets (Fig. 1).

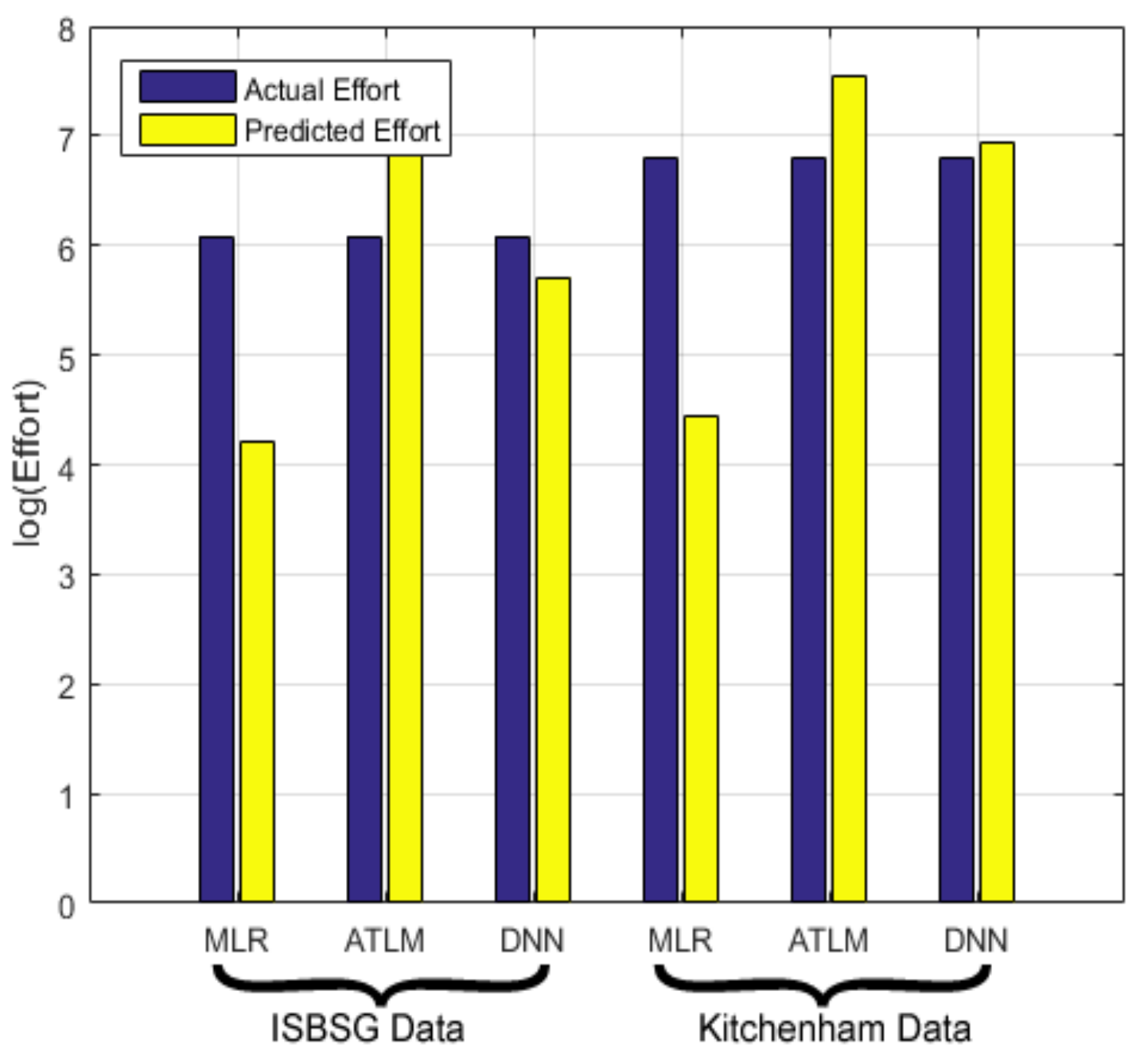

Fig. 1. New project's software effort estimation using the Bellwether moving window

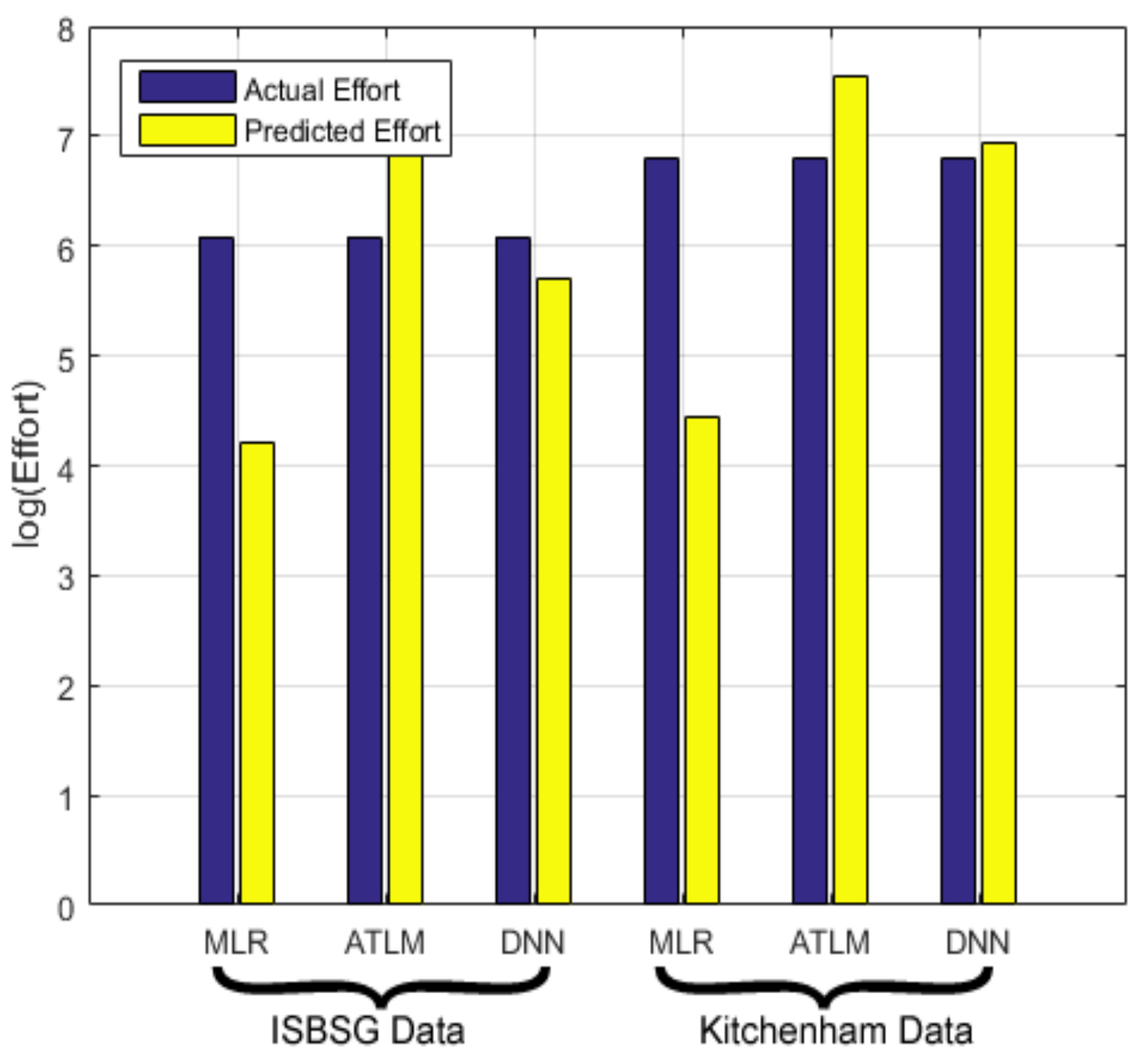

Fig. 1. New project's software effort estimation using the Bellwether moving window

Conclusion

The conclusion of the paper is that the C4.5 model provides better precision and higher recall for the top 30% of prospects when compared to the SBP model. The C4.5 model also provides a higher profit of $61,421,237 by choosing the top $3061,421,237 by choosing the top $3055,818,155 by choosing the top 30% of prospects. The authors suggest that the C4.5 model should be considered if one is looking for precision and higher recall, and that the SBP model may be more suitable if one is looking to achieve profits over $60,000,000 by choosing 40% of prospects.

Future Study

The paper suggests that the rules generated by the C4.5 and SBP algorithms could be further improved through the use of Genetic Algorithms, though the authors leave this as the topic of a further implementation study. The SBP Model and the C4.5 Model could also be combined to improve prediction.

There are several techniques available for combining models, such as Genetic Algorithms, Boosting, Stacking, and Collaborative Learning. Genetic Algorithms work well for combining knowledge/rules. The authors suggest that combining the rules from the models generated above using crossovers could lead to better rules.

Therefore, a further study could be conducted to explore the use of Genetic Algorithms and other techniques for combining models to improve the accuracy of the predictions

Referencaes

- P. A. Whigham, C. A. Owen, and S. G. Macdonell, "A Baseline Model for Software Effort Estimation," *ACM Trans. Softw. Eng. Methodol.*, vol. 24, no. 3, pp. 1–11, 2015.
- S. Amasaki and C. Lokan, "A replication study on the effects of weighted moving windows for software effort estimation," *Proc. 20th Int. Conf. Eval. Assess. Softw. Eng. - EASE '16*, pp. 1–9, 2016.
- R. Krishna, T. Menzies, and W. Fu, "Too much automation? the bellwether effect and its implications for transfer learning," *Proc. 31st IEEE/ACM Int. Conf. Autom. Softw. Eng. - ASE 2016*, pp. 122–131, 2016.
- R. P. Dobrow, "Introduction to Stochastic Processes With R," *Wiley Online Libr.*, pp. 181–222, 2016.
- C. Lokan and E. Mendes, "Investigating the use of moving windows to improve software effort prediction: a replicated study," *Empir. Softw. Eng.*, vol. 22, no. 2, pp. 716–767, 2016.