Advanced Machine Learning Kaggle - BNP Paribas Cardif Claims

Antonios Andronis[*], Pengfei GAO[†], Napoleon Koskinas[‡], Syed Muhammad Ali Shah[§] and Nikolaos Perrakis[¶]
University of Southampton
School of Electronics and Computer Science
Data Science Postgraduate Students
Email: [*]aa3e15@soton.ac.uk , [†]pg6g15@soton.ac.uk , [‡]nk5g15@soton.ac.uk , [§]smas1c15@soton.ac.uk , [¶]np4g15@soton.ac.uk

Abstract

The abstract goes here.

I. Introduction

Machine Learning is a field that merges Computer Science and Statistical Learning. It has been described as the “field of study which gives computers the ability to learn without being explicitly programmed”[1]. Of course computers have not yet become completely independent, but the evolution of Machine learning methods greatly contributed to the progress made in the filed of Artificial Intelligence. In recent years, when huge amounts of data concerning every sector are produced and stored worldwide, Machine Learning techniques have been flourishing as they enable computers to learn from available data and create models which enhance prediction and decision making. Data scientists, analysts and researchers develop and consequently apply these analytical modeling methods building on knowledge coming from pattern recognition and computational learning theory. In addition, they need to have a deep understanding of statistics, mathematical techniques and probabilistic modeling in order to apply the right method on every case.

In Machine Learning there is not a method which can be considered as ”panacea”, so the concerned Machine Learning practitioner should originally identify the kind of problem he has to face and analyze its nature to decide the process that needs to be applied. Firstly, he needs to note whether the problem belongs to supervised or unsupervised learning. Thereafter, comes the first contact with the available data which leads to the appropriate preprocessing techniques to be applied so that the data transforms into a format on which machine learning algorithms can perform. The actual analysis takes place when one or more machine learning methods are applied on training data and a prediction model is trained. This model is later applied on the test data in order to evaluate its efficiency on predicting unseen data by learning from data coming from the same source or distribution. In the end comes the evaluation of the model which can lead either to feedback to the whole analysis process in case the one applied did not work efficiently enough, or to the final conclusions and insights which can be extracted by modeling data.

The fore-mentioned authors, for the needs of Advanced Machine Learning module in the University of Southampton in academic year 2015-2016, decided to form a five-member group which undertook a real-world Machine Learning challenge and enabled them to apply the pipeline described above and obtain useful experience and skills.

II. Project Background

Insurance companies are working very hard to facilitate their clients for providing the best services especially at the time when their clients are facing sudden events in their life. In order to exceed the client expectation specifically in terms of claim approval and imbursements, insurance companies are using complex machine learning and data mining techniques to perform classification of true and false claims, risk assessment and payment process optimization.

False claims are increasing at the accelerated rate in UK[2] People are using very creative ideas about submitting false claims to insurance companies that apparently look a straightforward case for approval. It is a big challenge for the insurance company to timely classify the fraud claim and process the genuine case. Insurance companies uses the claim history and data from loss indicators to perform predictive analysis on people who can potentially do false claim.

The project team decided to research on insurance claim management subject by participating in Kaggle competition where a French insurance company named BNP Paribas Cardif has provided an opportunity to Data Scientists to work on the dataset the concerned company provided. Within this competition, Kagglers are challenged with assessing the validity of insurance claims.

The problem is a two-class classification problem, so the project team needed to apply supervised learning techniques in order to classify the claims as:

1) Claims which have enough information to process and decision could be taken straightaway for approval and payments.

2) Claims which have not enough information to process and need more information for decision.

The requirement of the competition from BNP Paribas Cardif was to predict a probability of classifying a claim to a class.

III. Dataset Exploration

IV. Preprocessing

V.

A. *Subsection Heading Here*

Subsection text here.

1) Subsubsection Heading Here: Subsubsection text here.

VI. Conclusion

The conclusion goes here.

Acknowledgment

The authors would like to thank...

References

[1] Phil Simon (March 18, 2013). *Too Big to Ignore: The Business Case for Big Data.* Wiley. p. 89. ISBN 978-1-118-63817-0.

[2] Kate Palmer (July 13, 2015) *Insurance cheats cost households 90 a year as bogus claims reach record high* http://www.telegraph.co.uk/journalists/kate-palmer/11736121/Insurance-cheats-cost-households-90-a-year-as-bogus-claims-reach-record-high.html