Introduction to Machine Learning and Predictive Analytics in Accounting, Finance and Management Minitrack

Peter Sarlin
Silo.AI
RiskLab at Arcada and Hanken School of Economics
peter@silo.ai

József Mezei
ÅboAkademiUniversity
Avaintec Oy, Helsinki
jmezei@abo.fi

The use of advanced statistical models, predictive analytics and machine learning have been present in the fields of accounting, finance and management for several decades. However, recent years have seen an ever increasingly growing trend of utilizing these approaches, as a result of the rapid evolution of related technologies mathematical algorithms. These developments include, but are not limited to: (i) the widespread availability of large amount of data, specifically data streams in new domains, (ii) the commoditization of advanced machine learning (ML) and artificial intelligence (AI) algorithms, such as deep learning in open source programming packages, (iii) the decrease in costs and complexity of performing computationally extensive modelling. These trends are observable in both academia and industry: for organizations, using AI and ML is not a source of competitive advantage anymore, but rather a necessity to remain profitable. The minitrack aims at showcasing some of the most interesting application domains and novel machine learning techniques applied to both structured and unstructured data sources.

This year we have received two times more submissions than in any of the previous years since the introduction of the minitrack in 2016. This indicates the increased general interest in and recognition of the importance of the covered topics. After the review process, six articles have been selected to be presented in the conference.

The first article "*An Explicative and Predictive Study of Employee Attrition using Tree-based Models*" is authored by Nesreen El-rayes, Michael Smith and Stephen Taylor (New Jersey Institute of Technology). The empirical study analyses data from the Glassdoor portal to understand the process of job transition processes and develop a predictive model for this purpose. After performing detailed descriptive analysis and feature engineering, several binary classification models are built. The authors find that tree-based models, such as random forests and light gradient boosted trees, offer the best prediction performance.

The second article "*Improving Credit Risk Analysis with Cluster Based Modeling and Threshold Selection*" by Ajay Byanjankar (Åbo Akademi University) investigates the performance of some machine learning models for credit scoring. The article combines segmented modeling with improved threshold selection and introduces the cost of misclassification in the confusion matrix. The approach is evaluated on a large real-life dataset from the Bondora peer-to-peer online lending platform. The results show that threshold selection performed independently for segmented models can improve classification performance, in particular when combined with random forests and gradient boosting trees.

In the next article, "*Enhancing Customer Satisfaction Analysis with a Machine Learning Approach: From a Perspective of Matching Customer Comment and Agent Note*", the focus is on analysing different types of user-generated content (UCG) in the context of customer services. The authors, Qiang Wei (Tsinghua University), Xiaowei Shi (Tsinghua University), Quan Li (BT China Research Centre) and Guoqing Chen (Tsinghua University) develop the CAMP approach, that combines customer comments, agent notes, matching analysis and net promoter score prediction. The approach involves a novel convolutional latent semantic model to analyse the two different UGCs and analyse semantic and sentiment in the text data. The developed tool can have important practical implications by improving customer follow-up services, and in turn increase customer satisfaction and customer loyalty.

In the fourth article entitled "*Towards Optimal Free Trade Agreement Utilization through Deep Learning Techniques*", Johannes Lahann, Martin Scheid and Peter Fettke (German Research Center for Artificial Intelligence, Saarland University) propose to introduce

URI: https://hdl.handle.net/10125/63912
978-0-9981331-3-3

deep learning models in the problem of Free Trade Agreement (FTA) analysis. As the authors state, prior literature and practice lack contributions on applying machine learning to generate insight from trade transactions to improve FTA utilization. In a multi-stage approach, the main focus is on identifying and making use of optimizable transactions that can aid in utilizing FTA more efficiently. The authors find that in this context deep learning techniques offer superior performance over more traditional machine learning techniques.

The fifth accepted submission, "*A CNN based system for predicting the implied volatility and option prices*", presents a study by Xiangyu Wei, Zhilong Xie, Rui Cheng and Qing Li (Southwestern University of Finance and Economics, The Key Laboratory of Financial Intelligence and Financial Engineering of Sichuan Province) focusing on issues related to the traditional problem of option valuation. As the authors observe, most of the existing models pose oftentimes unrealistic assumptions in the form of parametric models. In order to avoid this, a new deep neural network-based model is described to estimate implied volatility and combines the output of several parametric approaches to estimate the option price. The authors validate the model using data from the Chinese stock market observing a more than 40% increase in prediction performance.

In the final paper, "*Automating Lead Scoring with Machine Learning: An Experimental Study*", Robert Nygård (Idea Development BBN) and József Mezei (Åbo Akademi University) study how machine learning can be used to assist in analysing purchase data for the purpose of lead scoring. As the authors point out, while this is a highly relevant problem for every organization, academic research on predictive analytics for lead scoring is scarce. In the empirical analysis, purchase data from a business-to-business context is used to test the performance of some of the well-known classification algorithms. Additionally, different aggregation procedures are used to control for potential bias present in evaluating leads that were not converted into actual purchases. The authors find that random forest algorithm offers the best classification performance, and show how visual analytics can be applied to obtain novel business insights.