Ryerson University

Data Science
& Analytics
Graduate Studies

MRP Abstracts

2017

Abu-Ata, Muad Mustafa Husein - Optimization of Decision Model Microsimulation in Health Care

Microsimulation is used in health care to evaluate cost-effectiveness of different diagnosis and treatment procedures. Producing valid and statistically significant simulation results requires large input size. The aim of this project is to speed up an existing decision model simulation for Obstructive Sleep Apnea (OSA) study and to generalize the simulation to any diagnostic/treatment methods. Additionally, as mortality prediction is an important feature in such models, we aim to accurately incorporate mortality prediction into the simulation model. Parallelization and code refactoring are utilized to scale up the microsimulation model. We could scale up the simulation model to simulate four million patients in 21.8 minutes (reducing computational time by a factor of 14). Moreover, we applied the Lee-Carter model to future predict mortality rates where the fitted model resulted in small residual errors.

Chen, Yilin – New York City Green Taxi Trip Optimization

Most of people think driving a taxi is all about the driving skills, and there is no special rules or tricks to follow in order to let a taxi driver earn an outstanding amount of income.
How much a taxi driver could earn all depends on luck and long time working hours. However, what if there are some hidden tricks that could help a taxi driver to increase the daily revenue? This research paper aims to reveal those tricks and rules by digging into the big data world.
In this paper, I use the 2016 New York Green Taxi trip data from NYC open data source to generate an algorithm that takes expected starting location, the time of the day and date of the year as inputs, and outputs recommendations to taxi drivers on if the chosen expected starting location could earn the maximum revenue or the adjacent locations could earn higher revenues. The machine learning technique, random forest, is used to predict the factors that could affect the total revenue. The final simulation results indicate that by taking the recommendations provided by the algorithm, the revenue of a taxi driver is most likely to increase.

Durrani, Afsah – Filtering of Tweets to Identify and Remove Un-Informative Concepts

Due to the recent technological advancements, there is a large increase in the number of online users and the social media content generated. The abundance of online social media data is used by multiple stakeholders to identify the public opinions, trending topics and user segmentation. The large amount of data requires high computational power which is traditionally dealt with, by removing uninformative words using preprocessing techniques, such as stopword removal, before analysis. We present approaches using two correlation algorithms to identify the uninformative concepts. The effectiveness of the approach is evaluated by measuring the performance of the LDA models applied on the new datasets derived from the experiments. Correlation with the sum of all concepts performs better as compared to the correlation with the noise signal. Varying correlation threshold values are experimented with of which higher thresholds provide with better LDA performance.

Fatima, Hira - Analysis of Reddit Groups (Subreddits) Using Classification of Subreddit Posts

In this paper, we applied a novel idea to utilize machine learning techniques to automatically label subreddit posts from a subreddit called “askhistorians”. Using descriptive analytics, I first conducted an exploratory analysis to see if I can find any patterns, correlations or relationships that could be used to generalize posting pattern and behaviour of reddit users. The second part of my analysis comes from training and evaluating eight classifiers that could correctly categorize reddit posts with a positive or negative label for the eight category codes listed in Appendix A. I used 3 different algorithms and compared their performance using accuracy, precision and recall. This research is a continuation of an existing study that started in Ryerson Social Media Lab (RSML) [1]. The dataset that was used to train and evaluate the classifiers was coded manually by (Ryerson Social Media Lab) RSML. The predicted classification results were used to provide more insights about the subreddit group.

Ghaderi, Amir - Credit Card Fraud Detection Using Parallelized Bayesian Network Inferecing

The number of credit card transactions is growing, taking an ever-larger share of the worlds payment system. Improved credit card fraud detection techniques are required to maintain the viability of the worlds payment system. The aim of this Major Research project is (1) to develop a Bayesian network model that is able to predict fraudulent credit card transactions with minimal false positive predictions and (2) to reduce the processing time through the parallelization of the inferencing process. The Bayesian network was trained on credit card transaction data obtained from European cardholders for the month of September 2013. The results determined that Bayesian networks are able to be trained to predict fraudulent credit card transaction with zero false positive predictions. In addition, Bayesian network inferencing can be efficiently parallelized to reduce the overall processing time.

Ghaly, John - A Defect Prediction Model Using Delta Static Metrics

Dependence on software to automate, optimize and manage our daily tasks is growing every day. As the demand for higher software functionalities increase, the software size and complexity also increase. Maintaining and finding software defects are a hard and time-consuming job. We propose a machine learning model to identify and predict defect prone modules. We use an industrial dataset to build 8 classifiers from 5 different categories based on Static, Churn, and Delta metrics. We found that the addition of delta metrics significantly reduced the probability of false alarm while improving the probability of detection. Our results validate our hypothesis on the added value of delta metrics for improving results. We found that most algorithms achieved reasonable performance giving the suitable technique.

Hon, Marcia - Alzeheimer's Diagnosis with Convnet

Alzheimer's is a serious disease characterized by a progressive degeneration of the brain affecting 60 to 80 percent of dementia cases. The ability to automate this diagnosis is very important to accelerate treatment.
In this project, convolutional neural networks (convnets) are used in order to automate the classification of Alzheimer's disease from MRI images. 6400 MRI images were taken from http://www.oasis-brains.org using 5-fold with 80% to 20% test/validation. VGG16 won the ImageNet competition and it is thus used in this project. Its classification layer is retrained borrowing code from https://keras.io/applications/ and https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html using Keras, TensorFlow, and Python. A very high accuracy of 92% was achieved. This success proves that machine learning can successfully and readily be applied to medicine. Future projects could involve classifying skin cancer and other diseases with a visual component with different convnets. Additionally, if there is sufficient longitudinal data, MRI could be used to predict Alzheimer's instead of mere classification.

Islam, Md Shariful - Opinion Mining Classification of Twitter Data for Major Telecom Operators in Canada

A fierce competition is visible among the telecom operators to acquire more subscribers through advertisements and campaigns, especially in social media. Now the question arises how to measure the performance of operators based on customer response.
The goal of this project is to measure the competitive performance of mobile phone operators by analyzing the customer sentiment from twitter data and to build a classifier model by using different machine learning algorithms.
9000 tweets are collected for top three mobile operators in Canada. After data cleaning and text processing, sentiment analysis was completed. Then sentiments are classified and compared by using three different algorithms.
Among the three operators, Telus has the highest Positive sentiment than others. Among the algorithms, SVM and RF has better accuracy than decision tree.
This will help wireless operators to know about the negative experiences and to turn it into positive experience by improving the particular service.
.

Kundu, Somnath - Graph Theory Persepctive of Stock Market Behaviour

It is often noticed in practice that the prices of different Stocks and other financial investment instruments move together. It is not too surprising since many different

companies engage in similar type of business and one business depends on other businesses. So those companies are assumed to be tied together by an invisible thread of relationship. Though it is difficult to find the actual relationship of the companies we can always measure the strength of their relationship by the similarity of movement of their attributes.
We may assume that, if the correlation coefficient of one or more attributes of two stocks is larger than some chosen threshold value, then those two stocks are connected by an edge in the relationship graph.
In this project our objective is to explore these relationships between the stocks from the Graph Theory perspective and investigate various properties of this Stock Relationships Graph, including clustering.

Patel, Jitesh - Predicting Breast Cancer Survival Based on Gene Expression and Clinical Variables

Survival of breast cancer patients is irregular. Several gene sets are directly or indirectly involved in breast cancer. I explored whether the combination of mRNA expression of such sets may improve prediction of triple-negative breast cancer. I have used TCGA gene expression data for this study and classified 19 genes into two sets based on the relationship between death-risk and expression of each gene using Cox model. The up-regulation of the first gene-set combined with the down-regulation of the second gene-set is correlated with a high risk for the triple-negative breast cancer. The triple-negative breast cancer is classified based on the expressions of estrogen, progesterone and HER2 receptors. The combined effect of gene set1 and set2 on survival was predicted on overall data for triple-negative and luminal class of breast cancer using Kaplan-Meier model. Combining the effect of multiple gene signatures improves prediction of triple-negative breast cancer survival. This methodology can be relevant for different cancer types and target therapies.

Rizvi, Syed Ali Mutahir - Prediction of the Directional Change or Strength of Forex Rates

Predicting FOREX pairs direction or strength is extremely hard and predicting trends has been an area of interest for researchers for many years due to its complex and dynamic nature. There are hundreds of trend indicators for the prediction of FOREX but the accuracy is not reliable. In this project, a combination of indicators (Day Close Strategy, Moving Average Crossover, Fractal Strategy, Renko charts, ATR & Breakout) and machine learning algorithms (Naïve Bayes, Support Vector Machine, Deep Neural Networks) are used for better prediction accuracy and the results suggest that this approach is helpful in providing decision support.

Shi, Pengshuai - Population Counting with Convolutional Neural Networks

In this project, we explore the challenge of automatic population counting from single images. Most recent work apply Neural Networks to extract visual features and regress the population count either explicitly or implicitly. This type of model has been shown to perform better than traditional counting methods that require localizing each object and hand-crafted image feature representations. In this work, we compare two different types of CNN-based counting models. The first model consists of a fully Convolutional Neural Network (CNN) that predicts a pixel-wise density map, where the entity counts are realized by post-hoc summing over the density map. The second model is a neural network that consists of an initial set of convolutional layers, followed by fully connected layers that directly regresses the entity count. Our empirical evaluation considers three diverse data sets: (i) cells captured under a microscope, (ii) aerial views of sea lions, and (iii) aerial views of crowds of people. We find that the direct count regression approach generally performs better than the indirect one. In addition, we explore a saliency map approach to visualize the location of the count entities.

Trikha, Anil Kumar - Enhancing User Interest Representation in Social Media

User interest detection in social media is valuable for providing recommendations of goods and services, modeling users, and supporting online advertising. Only recently have models for inferring implicit interests and predicting future interests been proposed. We extend these models by specifying a technique that yields an improved representation of user interests for these purposes. We evaluate the solution on publicly available Twitter data. The research question we address is whether user interests derived from micro-blogging posts can be more accurately represented using a data mining approach that utilizes association rules.

Yadav, Shailendra Kadhka - Risk Prediction of Collisions in Toronto

Collision prediction models are used for a variety of purposes; most frequently to estimate the expected accident frequencies from various roadway entities like aggressive driving, traffic control, road class, speeding etc. and also to identify factors that are associated with the occurrence of accidents. In this study, the Decision Tree, Random Forest and ARIMA time series model are implemented and analyzed over the Killed or Seriously Injured (KSI) Traffic Data so as to predict the severity of injury type, number of collisions in Toronto for future 12 months. The ARIMA model gives accuracy of 85% for the prediction of number of collisions. The Decision Tree using CART and Radom Forest models returns accuracy of 57% and 67% respectively for the classification of injury types.

Yan, Bingsen - Automatic Sentiment Analysis Process: Amazon Online Shopping

Background: Sentiment Analysis appears to be significantly helpful offer of time-saving and efficiency enhancement especially given the fact that customers nowadays tend to rely more and more on products reviews when shopping online. Aim: Develop a Sentiment Analysis Tool as a Decision Aid for Online Shopping Experience Improvement. Methodology: We use web scraping technology to collect online real-time data, sentiment analysis technology to get sentiment score for each review and machine learning model to predict star rating. Results: We developed an automatic process to generate a product report including price, star rating, reviews and sentiment scores. Also, we analyzed the relationship between star rating and the three sentiment scores. In addition, a prediction model has been built up to predict star rating using sentiment score. Conclusion: The Tool enables customers to value a product in a more efficient way. Also, this is a powerful tool for star rating prediction.

Yueh, Ming-hui - Determining Factors Influencing Prediction of Length of Stay

Having a predictive model helps doctors identify short stay patients more objectively. To identify important features for predicting a patient's length of stay of 72 hours or less, three stages of data processing were performed from obtaining initial variables to applying feature selection methods for determining a subset of features, which were then fed into several learning algorithms. AUC and precision-recall curves were used to measure model performance. Regardless of the selection method and impute approach, ALB (Albumin) value, Age and HGB value were found to influence model performance the most.

2018

Amadou, Angelina – Geospatial Simulation and Modelling of Out-Of-Home Advertising Viewing Opportunity

Companies use out-of-home (OOH) advertising to promote their products. The purpose of the project is to build an integrated multi-source simulation model that allows Environics Analytics to establish optimal locations for marketing campaigns. The study area is located in the province of Manitoba in Canada. It concerns mainly the Winnipeg Metropolitan Area, which includes the city of Winnipeg and its surrounding municipalities. Using Dijkstra's algorithm for finding shortest paths, a simulation algorithm is developed. The top ten busiest intersections are retrieved and used as recommended locations for out-of-home (OOH) advertising. Additionally, a Wilcoxon signed rank test is used to validate the simulation output against empirical data. In general, there is no statistically significant difference between the simulated data and the empirical set. The study has shown that multi-agent-based models, although in their infancy, represent a viable approach to modelling population dynamic. Results from the simulation can be used to develop a new model which may include demographic profiles of the population for further studies.

Arabi, Aliasghar – Text Classification Using Deep Learning in Reddit Reply/Comments

In this paper, I implemented several deep learning models to automatically classify posts from subreddit of “askhistorians” into defined classes using pre-trained word embeddings vectors. The training data is taken from the research done at Ryerson University Social Media lab. I used one-vs-the-rest classifier (OvR) to train separate model for each of eight classes. Keras library from Python is used to develop deep learning frameworks starting from individual models such as CNN, and LSTM and finishing up by combining the individual models to form more complex versions such as CNN+LSTM and LSTM+CNN. When compared with previous work using traditional models and N-grams as features, improvement in all three accuracy, recall and precision is observed. However, the best model considering all evaluation metrics, stability/ranges of results for all iteration/fold, and run time found to be CNN for all categories.

Arjumand, Isra – Stock Market Prediction Using Machine Learning

This project focused to find efficient prediction of the Apple Inc. (AAPL) stock price movement to make effective investment decisions by generating trading decisions, comparing of SVM, KNN and RF machine learning algorithms and profit comparison. Research shows that use of machine learning algorithms with technical analysis gave good results. Technical analysis was implied on the data to generate trading signals and algorithms were trained on them to predict future stock trends. By applying trading rules, decision points (buy, sell and hold) are generated. SVM performed better on experiment 1 and RF was efficient in experiment 2. Performance was evaluated using profit percentage. Adding more technical indicators improved the profit percentage. In conclusion, better profits are generated when technical indicators were used along with machine learning techniques in contrast with technical indicators alone.

Beqaj, Inela – Diabetic Retinopathy Detection Using Convolutional Neural Networks

Machine learning techniques are becoming more and more helpful in many areas of our everyday life such as education, healthcare, etc. One of the main applications of these techniques in healthcare is computer-aided diagnosis which are systems that assist doctors in the interpretation of medical images. This project is focused on medical image analysis of retinal images to identify the type Diabetic Retinopathy eye disorder which is the leading cause of blindness among people diagnosed with diabetes. In this project are used supervised techniques and semi-supervised ones to classify the images. The two types of convolutional neural networks architectures applied in supervised learning are VGG16 and DenseNet121, while the architecture used in semi-supervised mode is Adversarial Autoencoder. The semi-supervised techniques achieve the same accuracy as the supervised one, but they are more efficient because they achieve the same accuracy using 10 percent of the labeled data

Chowdhury, Kakoli – Binary Classification on Clustered Data

Land-Mobile radio systems support many vital communication functions supporting government and private operators, some related to public safety and mission critical functions. The models produced will help in understanding the usage patterns at different time periods to predict occupancy and demand by different channels across the spectrum. CRC (Canadian Research Corporation) is providing Layer 1 data sampled every three milliseconds. This data is further explored and processed under this MRP. Sub-setting of data is conducted based on clustering and descriptive statistical analysis designed to differentiate between channels exhibiting different occupancy % patterns. Applying algorithms on clustered data is expected to show distinct behaviors that are further utilized to find the best prediction model for spectrum availability.

Fadel, Fady – Organizing Web Search Results Using Best Clusters Separation

In this paper, we applied a novel idea to utilize machine learning techniques to automatically organize web search results from search engine queries. Using text mining analytics, we first conducted analysis to identify features that can be used for clustering. Second part of analysis was an evaluation of the best clusters separation method and the performance comparison of the selected features against different clustering algorithms.

Gupta, Vasudev – Predicting Gold Prices Using Neural Networks

The aim of the study in this paper is to predict Gold future prices using neural networks. Prices of Gold change rapidly in real time across the globe, making the price prediction interesting and challenging. Predicting Gold prices stresses the machine learning algorithms and technology and is a good test case. The North American perspective on Gold price prediction was used within this study. Gold is used as an investment vehicle by large number of investors across the world and successful predictions can be very helpful. Five input variables were used to predict the price, which are: Silver Future price, Copper Future price, Dow Jones Industrial Average, US Dollar Index and VIX volatility Index. Two Types of Neural networks models were used to predict the Gold Prices: Feedforward Neural Network (FNN) and Recurrent Neural Network - Long Short-Term Memory (RNN-LSTM). As well, different variations of training data - weekly/daily, short/long term were tried. Experimentation was also undertaken with USD/CNH (US Dollar to Chinese Renminbi exchange rate) as an additional input variable.

He, Xin – Movie Recommender System: Using Ratings and Reviews

Because of information overload, it is becoming increasingly difficult for users to find the content that they are interested in. Usually, the actual ratings are used to implement a Recommender model. Currently, many item evaluation systems not only have the ratings but also the reviews. In this report, we mainly describe how to use both ratings and reviews to implement a recommender model. Additionally, the project investigates the relationship between the ratings and reviews.

Hyder, Md Khaled – Sentiment Analysis of Twitter Data For Top Canadian Retailers

The competition among retail companies is visible in all communication channels. Most companies are now focusing on social media marketing to reach the vast consumer. In parallel to aggressive communication retailers also want to measure their own and competitor’s performance.

This project aims to measure the performance of retail companies in Canada by analyzing users sentiment from tweets and build a machine learning model that can predict with higher accuracy, also conduct exploratory analysis to find user engagement and other hidden patterns.

286,668 tweets were collected for the top five retailers. After processing and cleaning the dataset, an exploratory analysis was conducted to find hidden patterns, and a sentiment classifier model developed using five algorithms experimented with two vectorizers.

Among the five retailers, Sobeys has the highest positive score than others. Initially, Linear SVM with count vectorizer produced the highest accuracy, then random oversampling with TF-IDF vectorizer produced high and balanced precision and recall values. This solution will help retailers to compare their performance with competitors.

Jain, Sachin – Binary Classification Prediction on Time Aggregated Data

The main objective of this Major Research Project (MRP) is to find out the effect of time resolution on the prediction of the channel occupancy of Land Mobile Radio channels to facilitate dynamic spectrum allocation that increases overall spectrum efficiency. This project is a collaboration between Canadian Research Corporation and Data science lab in Ryerson.

Layer 1 data is measured for occupancy percent of more than 7000 channels approximately every three milliseconds. This MRP specifically looks to generate aggregate dataset, generated from Layer 1 data, to predict channel occupancy. Further predictive classification is conducted using Naïve Bayes and Logistic Regression algorithms on the datasets. The ultimate goal of this project is to build spectrum occupancy prediction model that is known to work best in given conditions.

Jandu, Arshnoor – Neural Style Transfer With Image Super Resolution

In fine art, humans have mastered the skill to create unique visual representations through combining content and style of an image. However, rendering the semantic content of an image in different styles is a difficult image processing task. Recent success of Deep Learning in computer vision has demonstrated the power in creating imagery by technique of separating and recombining the image content and style called Neural Style Transfer (NST). Several online and offline optimization methods are proposed that produces new images of high perceptual quality. However, these existing methods do not offer flexibility of creating high resolution upscaled images. In this project, I have implemented deep neural networks for Neural Style Transfer and Single Image Super Resolution, where users can transform photos into desired paintings and further upscale them in high resolution quality. This project also demonstrates experimentation with several parameters of NST to create amazing photo effects.

Kashyap, Askhat – Stock Price Movement Prediction Using Social Media (Reddit) Analysis

In this paper, we applied different machine learning techniques to predict stock price movement based on metrics derived from reddit posts of a subreddit called "economy". As part of exploratory data analysis, I tried to identify patterns in stock price movement, performed data cleanup on reddit posts and identified important topics discussed in reddit posts. We categorized stock market data in 3 classes i.e. positive, negative and steady, we marked data as positive/negative if the market direction is upward/downward and more than a certain threshold (above +/- 1%) else we marked it as 'steady', we considered volume changes while calculating this percentage.

Khan, Ghazala – 6-Month Infant Brain MRI Segmentation With Convolutional Neural Network

Brain MRI segmentation and analysis is one of the most important and initial steps in measuring brain's anatomical structure and visualizing any changes and developments in the brain. Early stage of brain development is "critical in many neurodevelopmental and neuropsychiatric disorders, such as Schizophrenia and autism." These abnormalities and disorders are detectable at early infant age and early interventions are possible to control a life at risk.

To investigate the problem this paper proposed two models 2D Conv and 3D FCNN for the brain MRI tissues segmentation of 6 months infants into GM, WM and CSF with multi-modality T-1 and T-2 weighted images by using MICCAI grand challenge iSeg2017. The architecture of 2D Conv was inspired by VGG model with modifications. The architecture of 3D Fully Convolutional Neural Networks was inspired by the recent work on infant brain MRI segmentation.

The quantitative evaluation of 3D FCNN exhibited substantial advantages of the proposed method in terms of accuracy of tissue segmentation with efficient use of parameters. 3D FCNN has shown comparative performance with 21 state-of-the-art international teams of the iSeg2017 challenge and acquired DSC score of 93%.

Luo, Jiefan – Twitter Bots Detection Utilizing Multiple Machine Learning Algorithms

The purpose of this paper is to apply multiple machine learning algorithms to develop bot-detection models for Twitter. Using exploratory analysis, I explored the Twitter metadata and found useful behavior features to distinguish between normal users and bots. For the training models, I found optimal hyperparameters to tune the different models. I applied five algorithms including Naive Bayes, Decision Tree, Random Forest, Linear Support Vector Machine (SVM), and Radial Basis Function SVM to classify bots and humans. The results of the classification are the account identities, and I measured the classification performance by accuracy, sensitivity, specificity, and area under the receiver operating characteristics curve (AUC). The results present that the Random Forest algorithm was most effective in detecting bots and identifying normal users.

Najlis, Bernardo – Applications of Deep Learning and Parallel Processing Frameworks in Data Matching

Most of Data Science research work assumes a clean, deduplicated dataset as a pre-condition. In reality, 80% of the time spent in data science work is dedicated to data deduplication, cleanup and wrangling. Not enough research papers focus on data preparation and quality, even though it is one of the major issues to apply Data Science. The research subject of this paper is to improve Data Matching techniques on multiple datasets containing duplicate data using parallel programming and Deep Neural Networks. Parallel programming frameworks (like MapReduce, Apache Spark and Apache Beam) can dramatically increase the performance of computing pair comparisons to find potential duplicate record matches, due to O(n2) complexity of the problem. Deep Neural Networks have shown great results to improve accuracy on many traditional machine learning applications. The problem and solution researched are of general applicability to multiple data domains (healthcare, business).

Ong, Liza Robee – Predicting Depression Using Personality and Social Network Data

Over 300 million people worldwide suffer from depression. With the advent of social network, our goal is to apply a novel approach to identify depression by investigating what relationships exist between an individual’s social network information and speech features, their personality, and depression levels. The study was conducted using a publicly available dataset, called myPersonality, which contains more than 6,000,000 test results, and over 4,000,000 individual Facebook profiles. From the dataset, we used depression risk and personality assessment scores, Facebook network and linguistic measures. We created a classifier to extract a feature that indicates the speech act of a status update. We applied several machine learning methods and feature sets to predict depression risk based on personality type, speech acts, and network influence. Our results show that the best predictors included personality dimension scores on neuroticism, conscientiousness, and extraversion, and the usage scores for the assertive and expressive speech acts.

Rafayal, Syed – Tucker 2 Tensor Decomposition Model Implementation on Visual Dataset Using Tensor Factorization Toolbox

The main goal of this paper is to recognize and classify the images utilizing Tucker2 decomposition technique. In the first part of the experiment, the exploratory analysis is conducted. The second part includes building training models and automatically correctly label the testing images. In training and validation phase, different folds and values of indices (i, j) are used to have the best performance using accuracy. In addition, two approaches are adopted here for testing. The first approach randomly selects training samples from core tensors. In the second approach, the similarity score table is created and sorted in ascending order. The larger score of the image means more noise and core indexes are collected from every level of noise by a certain interval. Experimental results show the training models for the indices (i=8, j=8) obtained more success and the second testing approach is more consistent. All experiments have been conducted on a visual dataset which is a publicly available dataset called Fashion MNIST using MATLAB factorization software package known as Tensor Toolbox.

Rodrigue, Sami – Experiments With External Data and Non-Linearity for Channel Usage Prediction

Neural Networks are one of the most popular models for predicting channel usage in the telecommunication spectrum. They commonly use spectral, temporal or spatial information from simulated or cellular data. However, these sources can fail to capture the full array of user behavior. We will use fully connected Neural Networks and Perceptrons on LMR data collected in Ottawa to explore whether enriching the input space through the use of external data, such as weather data, applying non-linear transformation to the input space improves the predictive power of the models. Based on our initial analysis, we have failed to identify any improvement in prediction using weather data, however the benefit of non-linear transformation is dependent on channel behavior. The later point can be further explored via other models such as Recurrent Neural Networks and different grouping of the channels.

Sharma, Suansh – Spectrum Occupancy Prediction in Land Mobile Radio Using Multiple Hidden Markov Models

In this paper, we seek to predict the occupancy status of Land Mobile Radio channels based on real life spectrum measurements using machine learning techniques. Cognitive radios are essential for implementing dynamic spectrum sharing, which has been gaining attention as a promising solution to alleviate the problem of spectrum scarcity. HMMs are widely studied in the literature for spectrum prediction and by design, HMMs learn from the sequential nature of the data, which is directly applicable to case of temporal spectrum occupancy prediction. We implement a model made up of multiple HMMs to perform spectrum occupancy prediction. We use submodels to capture the primary user activity then the submodels are used to initialize a high-level HMM, which is trained over an LMR channel’s occupancy over the time. We validate the performance of the multiple Hidden Markov Models on LMR bands and show that the multiple HMM model performs better than single HMM on predicting occupancy status for the next hour. By training multiple HMMs, which capture not only channel occupancy patterns over time but also low-level user activity patterns, size-able gains can be made in the performance of data driven spectrum prediction techniques.

Sirwani, Naresh Kumar – Prediction of Query Hardness

Information retrieval (IR) became an important part of today's data driven world, although most IR systems suffers with high variance in their retrieval performance & results quality due to several reasons, even the system who performs better on average can still return poor results for some queries. Understanding such hard queries and in-fact predicting their difficulty level before the search is taken place can bring many improvements in performance of IR systems including but not limited to providing direct user feedback on expected quality of results, federation or metasearch, content enhancement and query expansion etc. In this paper, we systematically study & implement various TF-IDF based pre-retrieval method to determine queries difficulty level on different TREC's data collection, we then compare the results of our experiments with Neural embedding and SELM (Semantic Enabled Language Model) based models for which results are already available from other similar studies and find out which methods performs better, more relevant and accurate.

Taylor, Kisha – Automated Stock Trading Based on Predicted Direction of Next-Day Closing Prices – S&P 500 Index

This paper develops a model that tries to mimic a trader based on predicted direction of the next-day closing price of the S&P 500 ETF (Exchange Traded Fund) and can be applied to a single stock/index.

Three approaches are used:

(1) Technical analysis only

(2) Machine Learning (ML) using only closing prices as inputs (baseline models) and

(3) ML model ("hybridized inputs") that use a combination of technical indicator(s) and raw closing price as inputs.

This classification problem uses Accuracy (main metrics), Precision & Recall and return metrics. The data (sourced from Yahoo Finance) uses 3 ½ years of trading data (Open & Closing Prices) from 2-January-2015 to 06-June-2018.

The paper also explores the use of a buffer, examining its predictive impact. The buffer is essentially a threshold used to derive the signal generated by the technical indicator.

Tomini, Emmalie – Load Forecasting Using Recurrent Neural Networks in Ontario Energy Markets

Reliable electricity load forecasting is essential for industry to devise efficient energy management plans as well as in guiding conservation efforts. Rising market demand

and unpredictable behaviours has resulted in traditional methods of electricity prediction being no longer robust enough to accurately forecast market demand. The aim of this project was to use machine learning approaches to create a model for effective load forecasting. Implemented in python, a recurrent neural network was trained on a variety of input features in order to determine what information is necessary to model Ontario load patterns. Calendar variables such as day, month year, day of the week and time, as well as relative humidity and dew point temperature were determined to produce the most accurate results when the RNN model was trained on this input space, yielding a MAPE of 5.19% on the test set. The results obtained from the models implemented in this study produce reasonably accurate day ahead electricity forecasts. However, there is possibility for improvement in this field, and machine learning approaches provide an excellent application in this area of study.

Walia, Harneet – Customer Acquisition Through Direct Marketing Campaign Analysis

In this research, we analyze a direct marketing campaign dataset obtained from a Portuguese Financial institution to predict if a customer will subscriber to a fixed deposit(Upsell) along with predicting the month (time aware) to best reach out to the customer. To solve the issue of time-aware upselling, we have implemented Time Aware Upsell Prediction Framework (TAUPF) using two different approaches, with an aim to find the best approach and technique to build the prediction model.
TAUPF is implemented using Upsell Prediction Approach (UPA) and Clustered Upsell Prediction Approach (CUPA). We have also tried to answer the data imbalance problem by examining and comparing different methods of sampling (Up-sampling and down-sampling).
For decision tree, K-Nearest neighbor and Random forest it was observed that CUPA has higher F-Score than UPA. It is also observed that prediction of the month, the number of calls being made to the customer before the customer subscribes for a fixed deposit can be reduced by a significant number.

Wan, Alexander – Learning About Tensor Decomposition to Determine Length of Stay

Tensor decomposition is a section of data science that can be used to build prediction models. By using tensor decomposition on St. Michael’s Hospital data, a model can be developed to identify patient’s length of stay. An application of tensor decomposition known as generalized tensor product is applied on the St. Michael’s Hospital dataset. The dataset is assembled based on measuring variable’s performance, correlations for pre-processing data and keeping variables the hospital deems important. Using performance metrics like comparing other machine learning algorithms to measure model performance. The average error was around 80 hours from tensor decomposition. However, in comparison to the other machine learning algorithms, tensor decomposition was more accurate. A major problem was the computer used for this project was not powerful enough to test for higher dimensions. This could mean that the data needs to be looked at again to make a better dataset to analyze.

Wu, Xinjie – Validation and Sensitivity Study of a LSTM Model for Stock Price Prediction

Time series data is everywhere in everyday life as well as in many business sectors. Ability to predict the performance of a process in the future will help reduce uncertainty, risk, make the highest profit and best performance from many industries. Stock price sequence is an easy accessed on-going time series dataset. The “unpredictable” feature make it a good source to challenge emerging algorithms.

LSTM (Long Short Term Memory) is an algorithm well designed for time series data forecast. In this project, a recent proposed LSTM model for stock future price/movement prediction was studied and compared to other available models. Then the model was applied to couple selected stocks for validation. The sensitivity study of parameters in the model was also presented.

The study done showed the presented model had advantage over other models but are still not universal. Perfect prediction was not guaranteed.

2019

Afsar, Tazin – Chest X-Ray Segmentation Utilizing Convolutional Neural Network (CNN)

The proposal of this project is to analyze the Chest X-ray segmentation process using the improved Attention gate (AG) U-Net architecture. This model suppresses the irrelevant regions and saliently points out the useful features for the targeted relevant tasks. Also, it takes less computational resources and learns automatically for the different sizes and shapes of the target chest’s X-ray images. AG with U-Net increased the model’s sensitivity and accuracy. This experiment is a continuation of an existing work of “RSNA Pneumonia Detection Challenge” [27]. The proposed architecture is analyzed by using two renowned chest x-rays data-sets: Montgomery County and Shenzhen Hospital. The experimental result shows the improvement of dice scores and accuracy by 1.0% and 4.0% respectively compared with the existing standard U-net architecture.

Ahmed, Sayed – Effect of Dietary Patterns on Chronic Kidney Disease (CKD) Measures (ACR), and on the Mortality of CKD Patients

Chronic Kidney Disease (CKD) leading to End-Stage Renal Disease (ESRD) is very prevalent today. Over 37 millions of Americans have CKD. CKD/ESRD and interrelated diseases cause a majority of the early deaths. Many research studies have investigated the effects of drugs on CKD. However, less attention has been given to the study of the dietary patterns on CKD. This research study has uncovered significant correlations between dietary patterns and CKD mortality as well as identified diagnostic markers for CKD such as the Albumin to Creatinine Ratio (ACR). In this project, Dietary surveys from NHANES, and CKD Mortality dataset from USRDS were utilized to study the correlation between dietary patterns and morbidity of CKD patients. Principal Component Analysis and Regression were utilized to find the effect. Machine Learning Approaches including Regression, and Bayesian were applied to predict ACR values. Grains, Other Vegetables showed positive correlations with Mortality whereas Alcohol, Sugar, and Nuts showed negative correlations. ACR values were not found strongly correlated with dietary patterns. For ACR value prediction, 10 Fold Cross Validations with Polynomial Regression showed 95% accuracy.

Barolia, Imran – Synonym Detection with Knowledge Bases

This study presents distributed and pattern based approach to identify similar words in given tweets, using low level vector embedding in vector space model. To achieve good results using distributed approach, Bilinear scoring function has been calculated. Score (u,v) = $xxuuWW$xvt . $xxuu$ is the potential source word embedding and $xxvv$ are knowledge base seeds. Synonym seeds have been used from existing knowledge base (WordNet) and have been generated more synonyms that are not present in knowledge base but can be potential synonyms in given corpus. Term-Relevance Computational algorithm has also been used to identify synonyms that are specific to given corpus. Another approach that has been presented is pattern base approach. Co-occurrence matrix mas been prepared and it calculate the probability of occurring $xxuu$ and $xxvv$ within window size of 10. Low level embedding has been learned using conditional probability of $xxuu$ and $xxvv$. Result have been presented for both approach and best result has been achieved by combining both approaches. These approached have been evaluated by regenerating same and more synonyms from dataset and evaluated against existing knowledge base. Using distribution and pattern based approach with bilinear scoring function and conditional probability the precision and recall were 74% and 55% respectively which is quite good as other study find 60% precision and lower recall.

Boland, Daniel – Battery Dispatching for Peak Shaving Using Reinforcement Learning Approaches

Economic dispatch of energy resources such as batteries is an important and current problem. We apply three reinforcement learning algorithms, the Monte Carlo On-Policy and Off-Policy algorithms, and the DynaQ Planning algorithm, to a load-connected battery with time-of-use charges and a demand rate, to study the agent’s ability to converge towards a least-cost policy including peak-shaving. In two simple cases we use a fixed daily load profile, and in a third case we use 31-days of data to reflect uncertainty in the demand. In the simple cases, we observe the Monte Carlo agents converge more quickly and achieve better savings than the DynaQ agent, but all agents typically yield savings of only 40-50% of what is demonstrated to be possible after a

10,000-episode training time. The DynaQ agent significantly outperforms the Monte Carlo agents in the case of 31-days of data, highlighting planning behavior by reserving some charge and consistently achieving a higher degree of peak load reduction.

Cai, Yutian - Musculoskeletal Disorders Detection With Convolutional Neural Network

Musculoskeletal disorder is a common cause of chronic pain and movement impairment, which is diagnosed with medical imaging technologies such as X-rays. Due to the limited supply of skilled radiologists, the detection is expensive and time-consuming. In this project, we propose a model using machine learning or neural network techniques to perform the same task as radiologists in detecting abnormalities in Musculoskeletal X-ray. Musculoskeletal radiographs (MURA) is a large open-source radiograph image dataset that is used to develop and test our model. It contains labeled images for training and validation, as well as a hidden test to evaluate the model. Python will be employed in the project as it has a variety of package choices from statistical analysis to data visualization. We hope that the model can distinguish between the normal and abnormal X-ray studies and lead to significant advances in medical imaging technologies.

Choi, Claudia – Using Deep Learning and Satellite Imagery to Predict Road Safety

This paper expands on previous work combining satellite imagery and deep learning to predict road safety. Studies have shown support for the hypothesis that features of the built environment have an impact on city-scale issues and can be observed through satellite imagery. In this paper, a labelled dataset of satellite imagery was generated for the City of Toronto. Class balancing techniques were then used to mitigate model bias - the best technique was used for the experiments. A Convolution Neural Network (CNN) was trained for overall road accidents, pedestrian accidents and cyclist accidents. Each CNN model followed the ResNet50 framework pre-trained on ImageNet. The resulting high accuracy scores and low macro F1 scores indicate model sensitivity towards the majority class. The models were able to use observable features of the built environment to predict ‘highly safe’ regions but show poor performance on regions labelled has ‘highly risky’.

Chowdhury, Mushfique – Forecasting Sales and Return Products For Retail Corporation and Bridging Among Them

The purpose is to show how we can bridge between sales and return forecast data for each and every product of retail store by using the best model among several forecasting models & how management can utilize this information to improve customer satisfaction, inventory management or re-define policy for after sales support for specific products. The way of doing multi-product sales & return forecasting by choosing the best forecasting model (among several forecasting models) for every product was shown. Several machine learning algorithms has been used – ARIMA, Holt-Winters, STLF, Bagged Model, timetk, Prophet. For every product, best forecasting model was chosen after comparing all of these models to generate sales and return forecast data which was then used to classify every product as “Profitable”, “Risky” and “Neutral”. Experiment showed that 3% of total products was identified as “Risky” items in future. Management can use this information to take some crucial decisions. This paper showed how to compare different models to choose the best one for each and every product and dynamically choose the best model to generate sales and return forecast data without giving more focus in optimizing the models. This is completely a new approach of utilizing sales and return forecast data to give a unique insight to the management for taking informed decision for different crucial aspects as identified above.

Ensafi, Yasaman – Neural Network Approach For Seasonal Items Forecasting of a Retail Store

In recent years, there has been growing interest in the field of Neural Networks. However, for the task of seasonal time-series forecasting which has many real-world applications, different researches have shown varied results. In this paper, the performance of Neural Network methods in seasonal time-series forecasting has been compared with other methods. At first, classical timeseries forecasting methods like Seasonal ARIMA and Triple Exponential Smoothing have been used and then, more current methods like recently published model Prophet, Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN) have been applied. The dataset is public and consists of the sales history of retail store. The performance of different models has been compared to each other using different accuracy measurement methods such as RMSE and MAPE. The results showed the superiority of the Stacked LSTM over other methods and also, indicated the good performance of the Prophet model and CNN model.

Etwaroo, Rochelle - A Non-Factoid Question Answering System for Prior Art Search

A patent gives the owner of an invention the exclusive rights to make, use and sell their invention. Before a new patent application is filed, patent lawyers are required to engage in Prior Art Research to determine the likelihood that an invention is novel, valid or to make sense of the domain. To perform this search, existing platforms utilize keywords and Boolean Logic, which disregards the syntax and semantics of Natural Language and thus, making the search extremely difficult. Consequently, studies regarding semantics using neural embeddings exist, but these only consider a narrow number of unidirectional words. As a solution, we present a framework which considers bidirectional semantics, syntax and the thematic nature of natural language. The content of this paper is two-fold; BERT pre-trained embedding is used address the semantics and syntax of language, followed by the second component, which uses Topic Modelling to return a diverse combination of answers that covers all themes across domains.

Hosmani, Chaitra – User Interest Detection in Social Media Using Dynamic Link Prediction

Social media provides a platform for users to interact freely and share their opinions and ideas. Several researches have been conducted to predict user interests in social media. Because of the dynamic nature of social media, user interests change over time. In this paper, given a set of emerging topics and user's interest profile over these emerging topics we are interested to predict the user interest profile for the future. We conducted this experiment on twitter data captured for 2 months from 1 November 2011 to 1 January 2012. We will be using temporal latent space to infer characteristics of users and then predict user's future interests over these given topics. We will evaluate the results with different ranking metrics like MAP and nDCG. We will also compare our results with the results of Zhu et al. temporal latent space which uses the same methodology but on a different dataset.

Islam, Samiul – Product Backorder Prediction Using Machine Learning Techniques to Minimize Revenue Loss With Efficient Inventory Control

Prediction of backorders of products boosts up companies' revenues in many ways. In this work, we have predicted the backorder of products using two machine learning models named Distributed Random Forest (DRF) and Gradient Boosting Machine (GBM) in H2O platform and have compared their performances. We have observed that the GBM successfully identified approximately 94 products out of every 100 products those go on backorder. We have noticed that the current stock level and the lead time of products act as the deciding factor of products' backorder in approximately 45% of cases. We have shown how this model can be used to predict the probable backorder products before actual backorder can happen and visualize the impact on inventory management. Moreover, we have identified that the decision threshold below 0.3 for high probable backorder products and the threshold between 0.2 to 0.8 for low probable backorder products maximizes organizational profit.

House-Senapati, Kristie - The Use of Recommender Systems for Defect Rediscoveries

Software defects are a known issue in the world of technological advancement. Software defects lead to the disruption of services for a customer, which in turn results in customer dissatisfaction. It is not feasible for all customers to install a fix for every known defect as this requires extra resources. Our goal is to predict which future defects a customer may discover, so that a fix can be put into place before the customer discovers that defect. We use recommender systems to build a predictive model. We evaluate our approach with publicly available datasets mined from Bugzilla (Apache, Eclipse and KDE). The datasets contain information about approximately 914,000 defects over a period of 18 years. From our experiments, we find that the popular algorithm performs the best with average Matthew Correlation Coefficient of 0.051. We also observe that the Funk SVD, apriori, eclat and random algorithm perform poorly.

Husna, Asma - Demand Forecasting in Supply Chain Management Using Different Deep Learning Methods

Supply Chain Management (SCM) is a very fast growing and largely studied field of research that is gaining in popularity and importance. Most organizations focus on cost optimization and maintaining optimum inventory levels for consumer satisfaction, where Machine Learning techniques helps these companies. The main goal of this paper is to forecast the unit sales of thousands of items sold at different chain stores located in Ecuador. Three deep learning approaches Artificial Neural Network (ANN), Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) are adopted here for better predictions from the Corporación Favorita Grocery Sales Forecasting dataset collected from Kaggle website. Finally, the performances are evaluated and compared. The results show the LSTM network tends to outperform the other two approaches in terms of performance. All experiments have been conducted using Python's deep learning library and Keras and Tensorflow packages.

Lee, Veson – Estimating Volatility Using A LSTM Hybrid Neural Network

Volatility estimates of market traded financial instruments are used in risk management models and portfolio selection. Hybrid neural networks are neural networks which combine a traditional parametric model such as GJR-GARCH with a neural network component and have been shown to improve volatility predictions. This paper will examine hybrid neural networks incorporating two different neural architectures, one with a LSTM, without exogenous explanatory variables and measure its performance using data from the Toronto Stock Exchange and S&P 500. We found that, in a neural network without exogenous explanatory variables, hybridizing the network by incorporating the parameters from a GJRGARCH(1,1,1) model does seem to have some possible benefits.

Matta, Rafik - Deep Learning to Enhance Momentum Trading Strategies using LSTM on the Canadian Stock Market

Applying machine learning techniques to historical stock market data has recently gained traction, mostly focusing on the American stock market. We add to the literature by applying similar methods to the Canadian stock market, focusing on time series analysis for basic momentum as a starting point. We apply long-short term memory networks (LSTMs), a type of recurrent neural network and do a comparative analysis of the results of a LSTM to a logistic regression (LOG) approach as well as a basic momentum strategy for portfolio formation. Our results show that the LSTM financially outperforms both the LOG and basic momentum strategy, however the area under the curve of the receiver operating characteristic curves show the results do not outperform a random walk selection. We conclude that there might not be enough data in monthly returns for the LSTM in its current configuration.

Natarajan, Rajaram - Road Networks – Intersections and Traffic

For this MRP, Research will be conducted on traffic/congestion in the Road Network. Our goal was to identify the most critical intersection, looking for ways to improve traffic and what can make the traffic worse. I will build a simulation model, research the city road network and the simulated traffic, I will show the areas which has the highest and lowest congestion based on the simulation, and which areas has the most efficient flows in traffic. I will also evaluate the impact on traffic when a critical node is brought down, impact on high congestion area and in the overall network. I will talk about the connections and relationships between the most critical nodes. Based on the research, recommendation will be provided on the changes in the road including new road/bridge construction that reduce the traffic and ways to reduce congestion. I will be using the Ottawa-Gatineau Dataset and Julia language.

Ozyegen, Ozan – Experimental Results on the Impact of Memory in Neural Networks for Spectrum Prediction in Land Mobile Radio Bands

Land-Mobile radio systems support many vital communication functions supporting government and private operators, some related to public safety and mission critical functions. The models produced will help in understanding the usage patterns at different time periods to predict occupancy and demand by different channels across the spectrum. CRC (Canadian Research Corporation) is providing Layer 3 data sampled every hour. This data is further explored and processed under this MRP. A powerful learning algorithm called Long Short Term Memory Networks is used to predict the occupancy of LMR bands over multiple time horizons. The results are compared with a seasonal ARIMA model and a Time Delay Neural Network. Results show that LSTM prediction models that remember long term dependencies and thus designed to work with time series data provide a better alternative for accurately predicting spectrum occupancy in bands that exhibit similar characteristics to LMR channels, especially as the forecast horizon gets longer.

Patel, Eisha – Generating Stylistic Images Using Convolutional Neural Networks

Fine arts have long been considered a reserved mastery for the minority of talented individuals in society. The ability to create paintings using unique visual components such as color, stroke, theme, etc. is currently beyond the reach of computer algorithms. However, there exist algorithms which have the capability of imitating an artist’s painting style and stamping it on to virtually any image to create a one-of-a-kind piece. This paper introduces the concept of using a convolution neural network (ConvNet or CNN) to individually separate and recombine the style and content of arbitrary images to generate perceptually striking “*art*” [2]. Given a content and style image as reference, a pre-trained VGG-16 ConvNet can extract feature maps from various layers. Feature maps hold semantic information about both reference images. Loss functions can be developed for content and style by minimizing the mean-square-error between the feature maps used. These loss functions can be additively combined and optimized to render a stylistic image [6]. This technique is called Neural Style Transfer (NST) and it was originally developed by Leon Gatys in his 2015 research paper, "*A Neural Algorithm of Artistic Style*". My MRP research attempts to replicate and improve upon the work done by Leon Gatys. The purpose of this research is to experiment using a variety of feature maps and tweaking the loss function to identify visually appealing results. A total variation loss factor is also included to minimize pixilation and sharpen feature formation. Images generated have been assigned a Mean Opinion Score (MOS) by a group of non-bias individuals to affirm the attractiveness of the results.

Peachey Higdon, Ben – Time-Series-Based Classification of Financial Forecasting Discrepancies

We aim to classify financial discrepancies between actual and forecasted performance into categories of commentaries that an analyst would write when describing the variation. We propose analyzing financial time series leading up to the discrepancy in order to perform the classification. We investigate what models are best suited towards this problem. Two simple time series classification algorithms – 1-nearest neighbour with dynamic time warping (1-NN DTW) and time series forest – and long short-term memory (LSTM) networks are compared to common machine learning algorithms. We perform our experiments for two cases: binary and multiclass classification. We examine the effect of including supporting datasets such as customer sales data and inventory. We also consider augmenting the data with noise as an alternative to random oversampling. The LSTM and 1-NN DTW models are found to be the strongest, suggesting that the time series approach is appropriate. Including the inventory dataset improves multiclass classification. Data augmentation grants a slight improvement for some models over random oversampling.

Postma, Cassandra - Netflix Movie Recommendation Using Hybrid Learning Algorithms and Link Prediction

Netflix, a streaming service that allows customers to watch a wide variety of movies, is constantly updating and optimizing their search and recommendation algorithms to improve user experience. The aim of this paper is to recommend movies to users using different link prediction methods as well as predict a user's movie rating using a comparison of various learning algorithms. First, an exploratory analysis was conducted to find any correlations between variables and users. Then several algorithms were used to predict a user's movie rating such as KNN, SVM, and hybrid learning algorithms. Finally, the data was represented as a graph and several link prediction algorithms were run to compare different recommender systems.

Ragbeer, Julien – Peak Tracker

IESO is the Crown Corporation responsible for operating the electricity market in the province of Ontario, Canada. IESO publishes (every-changing) forecasts for what it expects Ontario electrical demand to be in the near future. In this paper, we focus on short term time-series forecasting (within 24 hours). This solution hopes to forecast better than IESO, so that large commercial customers can feel surer about what the upcoming demand is, and when to shave power if they are Class A customers. The solution combines many data sources (weather forecast data, historical weather data and historical demand data) and aggregates them. The project uses numerous regressors (both linear and non-linear) on the aggregated data to come up with a prediction which is compared to IESO's forecast using 3 metrics, coefficient of determination, mean absolute error and the number of times it correctly predicts the hour of the highest daily demand. The results of this paper (10%-40% more accurate than IESO in some cases) show that there's value in out-predicting IESO's free model – being more accurate can have a positive effect on the bottom-line.

Raja, Abdur Rehman – Rating Prediction of Movielens Dataset

In the modern world convenience has become the biggest factor in our modern-day lives. Due to the overwhelming choices each consumer has there is a need to filter, prioritize and efficiently deliver recommendations to consumers. This project aims to look at one of the most famous datasets provided by GroupLens for research purposes called MovieLens 20M. GroupLens is a research lab trying to advance the theory and practice of social computing. GroupLens has collected and made available rating datasets from the MovieLens website which is a free movie recommendation service. The project will look at finding the best solution to predict the movie ratings to be used in a recommender system algorithm. One of the main algorithms we use that is discussed in detail is BellKor's Solution which is the algorithm the winner used in the Netflix competition to predict movie ratings. A comparison of BellKor's solution and other algorithms take place to find the best algorithm suited for this dataset.

Roginsky, Sophie – Radio Coverage Prediction in Urban Environment Using Regression-Based Machine Learning Models

Having a reliable predictive model of radio signal strength is an essential tool for planning and designing a radio network. The propagation model is often used to determine the optimal location of radio transmitters in order to optimize the power coverage in a geographic area of interest. This research proposes a Generalized Linear Model for radio signal strength prediction. Using feature engineering methods, the performance of the linear model was optimized to offer predictive accuracy comparable to more complex regression models, i.e. Multi-Layer Perceptrons and Support Vector Regressors, found in existing literature. Beyond computational efficiency, the advantage of the GLM is that it is linear in parameters, making it a viable option for coverage optimization applications.

Saeed, Usman - Digital Text Forensic: Bots and Gender Classification on Twitter Data

This research work describes the contribution of the Data Science department of Ryerson University, Canada in task bots and gender profiling at CLEF PAN-191 evaluation lab. The goal of this paper is to detect (A) if the author of a Tweet is a bot or a human, (B) if human, identify the gender of that particular author. Data set was made available by PAN lab. We participated in the research of English language data set only. In the proposed approach, before applying machine learning models, we used different word vectorization techniques after applying various preprocessing techniques (stemming, stop words removal, lowercase, etc.) on the data set. On independent evaluation of PAN lab test dataset, we got best accuracies 79.51 on task A (binary class) by using MultinomialNB and 56.55 on task B (multi-class) by using Decision Tree classifier.

Sokalska, Iwona - Boosting Bug Localization with Visual Input and Self-Attention

Deep Learning (DL) methods have been shown to achieve higher Mean Reciprocal Rank (MRR) scores in bug-localization compared to Information Retrieval (IR) methods alone. A combination of both can boost scores by 6% for MRR of 48%. The DL model consists of Recurrent Neural Network. In natural language research, it has been demonstrated that RNN neural networks with visual input and 'attention mechanism' are more robust at tasks that require incorporation of distant information. The objective is to examine whether an RNN network with attention mechanism using images of code snippets can achieve higher scores than an RNN alone. Moreover, to see if the improved performance is in a similar range as the improved performance between standalone RNN vs. RNN + IR. Using the data gathered from the open source Spring-Boot project, covering data from 2013-2018, a baseline RNN model was compared to an enhanced model RNN with a supportive convolutional neural network that analyses the image of the source code. A 5-fold experiment was conducted to compare the baseline model with the 2 test models. Two test models differ only with the usage of self-attention in the convolutional branch. The test model with self-attention had the highest mean accuracy across the 5 folds, of 61.98 in comparison to 60.70 of the base model. The two-tailed Welch t-test reveals that this difference between the means is not statistically significant. In contrast the IR methods on average provided 6% boost to the scores.

Tabassum, Anika – Developing a Confidence Measure Based Evaluation Metric for Breast Cancer Screening Using Bayesian Neural Networks

Screening mammograms is the gold standard for detecting breast cancer early. While a good amount of work has been performed on mammography image classification and many of the recent ones have made use of deep neural networks successfully, there has not been much exploration into the confidence or uncertainty measurement of the classification, especially with Bayesian neural networks. In this paper, we propose a new evaluation criterion based on confidence measurement for breast cancer mammography image classification, so that in addition to classification accuracy, it provides a few numeric parameters that can be tuned to adjust the confidence level of the classification. We demonstrate the use of Bayesian neural networks and transfer learning in the process of achieving that. We also demonstrate the expected behaviour resulting from tuning of the parameters and conclude by saying that the approach is extendable to any domain in general and any number of classes.

Zhang, Shulin – Artificial Neural Networks in Modelling the Term Structure of Interest Rates

In this paper, we applied Artificial Neural Network (ANN) to model the term structure of interest rates. In the exploratory analysis we observed the trend of the yield curve since 1991 to understand the underlying pattern. The Principle Component Analysis is employed to construct the input dataset as well as serving as a baseline model. We used different hyper-parameters, customized loss function and implemented regularization to tune the ANN model. The result section discussed the selection of best model and the prediction differences between PCA and ANN. ANN can match PCA results in a very limited case of strong regularization. The ANN has the potential to replace PCA but a careful design needs to be reviewed. This project is a continuation of an existing study that Dr. Alexey Rubtsov started for Global Risk Institute. The predicted analysis is used to provide more insights on financial applications of ANNs.

Zhao, Xin – Station Based Bike Sharing Demand Prediction

Bike-sharing have been increasing popularity in recent years due to its usage flexibility, reduction in traffic congestion and carbon footprint. Being able to accurately predict each bike-sharing station’s demand at any given hour is crucial for inventory management. This report first manipulated Bike Share Toronto ridership data with Toronto City Center weather data from 2016 Quarter 4 to 2017 Quarter 4, then implemented machine learning algorithms in particular Regression Trees, Random Forest, and Gradient Boosting Machine (GBM) to forecast station based hourly bike-sharing demand in the City of Toronto. The results indicated that Random Forest based prediction model was the most accurate model by comparing Root Mean Square Error (RMSE) of all bike-sharing stations.