Group 4 Literary Review

By: Briana Olsen, Charlie Simon, Sarah Sullivan and Pavin Steele

Listed below are brief summaries of 4 literary reviews that each group member read. After all summaries have been posted, a summary of these literary reviews gives the insight and takeaways that we learned about the specific scientific journals that we read.

Briana:

**Application of Deep Learning for Credit Card Approval: A Comparison with Two Machine Learning Techniques**

Using a deep learning model and two machine learning algorithms for credit card risk management.

Intro: Deep learning is a most popular and accurate classification technique that outperforms other machine learning models (e.g. logistic regression (LR), linear discriminant analysis (LDA), multiple discriminant analysis (MDA), k-nearest neighbor (k-NN), decision trees, etc.)

Models:

- Logistic Regression: One of the most commonly applied statistical techniques for credit card analysis. It predicts the likelihood of a result that can just have two states (i.e. a dichotomy). The prediction depends on the use of one or several indicators (numerical and categorical).

- Support Vector Machine: An algorithm that learns based on instances given and predicts. It is used as a classification and regression tool to maximize predictive accuracy. An SVM can learn to recognize fraudulent credit card activity by examining hundreds or thousands of fraudulent and non fraudulent credit card activity reports.

- Deep Learning: A subset of machine learning methods based on artificial neural networks. Core concept of deep learning is automating the extraction of features from the data. “Deep learning is a class of machine learning algorithms that: (1) use a cascade of multiple layers of nonlinear processing units for feature extraction and transformation. Each successive layer uses the output from the previous layer as input, (2) learn multiple levels of representations that correspond to different levels of abstraction; the levels form a hierarchy of concepts.”

Paper goes over Data Pre-processing, Data Analyzing, Experimental Design and Evaluation.

These models are worth looking into and testing/comparing. Great resource.

https://www.researchgate.net/profile/Md-Kibria-12/publication/348755769\_Application\_of\_Deep\_Learning\_for\_Credit\_Card\_Approval\_A\_Comparison\_with\_Two\_Machine\_Learning\_Techniques/links/600f495f92851c13fe39bb38/Application-of-Deep-Learning-for-Credit-Card-Approval-A-Comparison-with-Two-Machine-Learning-Techniques.

**The Application of Machine Learning Algorithms in Credit Card Default Predic**

Results for this paper: Random Forest model returns 82.12% accuracy.

Models:

- Logistic Regression

- Decision Tree: A method that can address nonlinear problems.

- Random Forest: A traditional ensemble learning method, for regression and classification. (Drawbacks such as low stability and accuracy can be avoided using ensemble learning to predict default possibility.)

If we want to look into decision trees and random forests, this may be a good resource.

<https://ieeexplore.ieee.org/abstract/document/9275986?casa_token=bXaSQU0MnRQAAAAA:goSh_BnCokm7G3-Tscif7ssxwnEksOIm4XpdKYnPPdNPSddEopmNsv4LC4htTMA8Sr3N2IDT0Z0>

Charlie:

**Credit Card Approval Prediction using Classification Algorithms**

During preprocessing the authors performed min-max scaling, one-hot encoding and feature engineering to address the skewness of the data. The authors are not specific about how or what they did during this phase. For Exploratory data analysis, ANOVA, correlation and chi-squared analyses were performed to determine relations between variables. The authors specifically mention an ANOVA analysis between job title and age but do not give further details.

The authors used 10 different classification models and spoke briefly on those that were valuable. The authors state that accuracy cannot be considered a reliable parameter. Precision and recall are more viable; however, they cannot be optimized at the same time, so F1 score is used because it balances the two. SVM, Adaboost Classifier and Gradient Boosting Classifiers performed the best of all algorithms. The following table summarizes the performance of each method:

A table with numbers and text

Description automatically generated

The authors state that no attempt to rebalance the data after defining the classifications was made. They state that the rules should be dependent on whether the market is a bull or bear market. Importantly the division between approved and not improved is much closer to 50:50 than our criteria, so imbalanced classification is not as much of an issue.

Dalsania, N., Punatar, D., & Kothari, D. (2022). Credit card approval prediction using classification algorithms. *International Journal for Research in Applied Science and Engineering Technology*, *10*(11), 507–514. https://doi.org/10.22214/ijraset.2022.47369

Sarah:

**Credit Card Ascent Using Supervised Learning** A screenshot of a computer

Description automatically generated

Credit Card Ascent Using Supervised Learning focuses on using supervised learning techniques seek to eliminate risk in credit card approval. Manually approving credit cards can take a lot of time, so applying an effective algorithm could save companies lots of time. This study uses many similar metrics of information based on an application, such as age and gender. It utilized KNN, Decision Trees, Naive Bayes, and Logistic Regression. The study concludes that a Decision Tree was the most accurate for judging credit card approval.

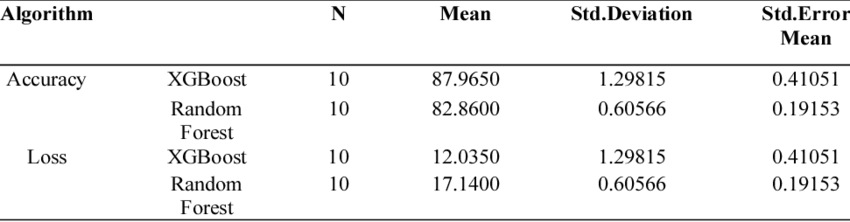
Pavin:

# Improve Accuracy in Prediction of Credit Card Approval Using Novel XGboost Compared with Random Forest

A compilation of different algorithms was used to determine the highest accuracy for an approval rate model and came up with the highest accuracy belonging to decision tree algorithms with a 89% accuracy. Because a decision tree algorithm was determined to have the highest accuracy, this study aimed at trying the random forest algorithm, a variation of the decision tree algorithm, to see if accuracy could be improved. Random forest algorithm considers multiple decision trees, making the accuracy significantly higher with less overfitting.

Another algorithm that was tested to see the accuracy of is Xgboost Classifier. Xgboost Classifier is a scikit-learn machine learning formula that uses wrapped classes and splits the entire data frame into smaller subframes. These subframes are randomized and run through the training model to try and produce a formula with a low threshold, high accuracy and low loss.

After testing both formulas with a credit card dataset from Kaggle, the random forest algorithm came back with a 82% accuracy while the Xgboost Classifier came back with a mean accuracy of 87%. This shows that within this certain data set, Xgboost is the best algorithm for determining the approval rate of a credit card dataset.



<https://www.researchgate.net/publication/365078059_Improve_Accuracy_in_Prediction_of_Credit_Card_Approval_Using_Novel_XGboost_Compared_with_Random_Forest>

Summary:

These different scientific articles help us decide which algorithms we should use to get the highest accuracy in relation to our data set and the decision process of which variables to use. Most of these articles talked extensively about Decision Trees and 2 of them recommended us using them because they resulted in the highest accuracy. Briana’s and Sarah’s articles gave high praise to the use of Decision Trees, meaning that this algorithm is something we should highly consider in our own data set decisioning. Pavin’s literary review focussed on different variations of Decision Trees and if we could further improve on the accuracy of this algorithm.

Charlie’s article focused on another type of classification algorithms like Adaboost Classifier, Support Vector Classifier and Gradient Boosting Algorithm. The Gradient Boosting Algorithm resulted in the highest accuracy. However, the study mentions that accuracy might not always be the best metric to use so precision and recall are more appropriate. Of which, Gradient Boosting Algorithm also scored highest out of all three classification algorithms.

Taking into consideration all four of these articles, we will seriously consider using decision trees and some variation of a classification algorithm (more than likely Gradient Boosting Algorithm) to decide which variables best fit our prediction model. These articles continue to provide helpful insight into which metrics and possible variations of algorithms we must consider when deciding on useful decision algorithms.