Explainable Machine Learning: Credit Scoring for Home Equity Borrowers

Machine Learning Engineer Nanodegree

Capstone Proposal

Shivraj Bheenick

June 3,2018

1. Domain Background

We are currently witnessing an unprecedented surge in the adoption of machine learning systems, with applications ranging from autonomous cars to ground-breaking medical diagnosis. As these intelligent systems become more prevalent in today's world, we will become ever more reliant on the decisions and recommendations provided by machines in our daily lives. While research in artificial intelligence continues to enhance the cognitive capabilities of machines, the outcome of this budding human-computer partnership will largely depend on a fundamental value that underpins any relationship: trust.

Current machine learning models are akin to black-box models and their interaction with human experts offers limited transparency on the rationale for the predictions or recommendations which they generate. This issue is exacerbated by the fact that predictive models tend to favour statistical correlations and significance over causality. Consequently, concerns on the fairness and ethics of artificial intelligence systems should be addressed to build credibility with human users. The EU's recent GPDR regulations contain a "right to explanation" clause where users are entitled to ask for the reasoning of an algorithmic decision made about them, further highlighting the importance of trustworthy models.

According to the XAI concept (Figure 1) by DARPA, the next generation of machine learning systems will be designed to have an explainable model and interface to enable human experts to understand the rationale behind the decision making and identify the strengths and limitations of the model. Designed to be human interpretable, explainable machine learning systems can be assessed and enhanced to ensure fairness, reliability, privacy protection and causality – all key factors contributing to building human trust in the automated decision-maker.

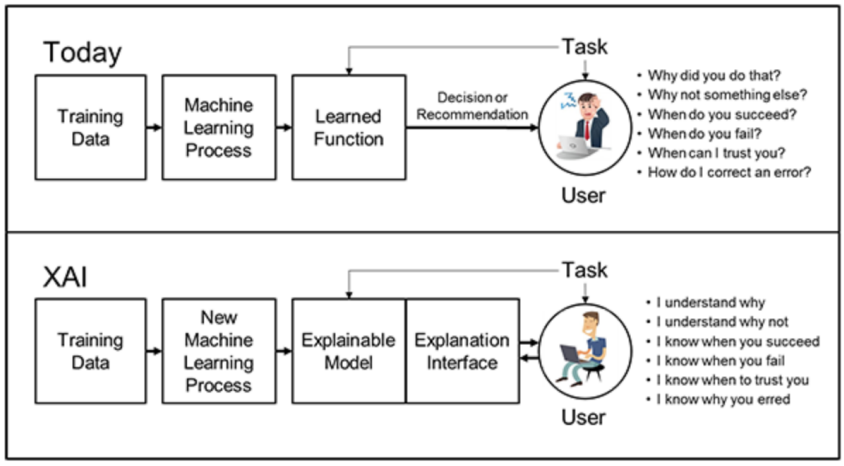

Figure 1: XAI systems will have an explainable model and an interface to interact with human users

2. Problem Statement

Credit scores are used to assess consumer credit risk in terms of the likelihood of repaying a borrowed amount as per the covenants agreed. Lenders use the credit scores to assess how much credit should be granted to a borrower, if at all, and what are the terms and interest rates for any credit extended. As sophisticated machine learning techniques increase the accuracy of credit scoring, there is an equally pressing need for greater explainability of these models to regulators, consumers as well as credit analysts. While regulators are keen to supervise scoring practices to ensure fairness, consumers are entitled to know the basis for their individual credit score and how this can be improved. Finally, explainability will also help credit analysts understand their datasets and the models' predictions, detect and correct for biases, and ultimately create better models.

As part of the FICO Explainable Machine Learning Challenge, machine learning techniques will be used to assess the credit risk of home equity borrowers who are seeking a Home Equity Line of Credit (HELOC)[1] in the range of $5,000 - $150,000. Each borrower will be assigned a binary RiskPerformance classification. A "Bad" RiskPerformance indicates that a consumer was 90 days past due or worse at least once over a period of 24 months from when the credit account was opened, while a "Good" value implies the borrower has made his payments without ever being more than 90 days overdue. This classification will help determine whether the homeowner qualifies for a line of credit and, if so, how much credit should be extended.

The two key objectives of this project are to maximise the prediction performance of the model while retaining the explainability of the model. To meet the latter requirement, two types of explanations will be generated. Firstly, a global explanation will provide transparency over the overall prediction mechanism of the model, including which features are the most important in determining the credit score. A local explanation will also be generated for each customer prediction to justify individual credit scores and identify how the customer can improve his credit score to have a better chance of getting a loan in the future.

3. Datasets and Inputs

FICO has provided an anonymised dataset of roughly 10 thousand HELOC applications made by real homeowners. The dataset has 23 features which relate to different aspects of creditworthiness such as length of credit, payment history and amount of debt. The target variable is equally represented in the dataset, so there are no imbalances to consider.

A data dictionary file has also been provided to explain each feature and its relationship to the target variable, which is captured by the monotonicity constraint. All the features are either quantitative or categorical, while the target variable (RiskPerformance) is binary: Good or Bad.

4. Solution Statement

Supervised machine learning classification techniques will be used to predict the RiskPerformance with a high degree of transparency and prediction performance. 4 models will be explored in this project:

1. Logistic regression – this will be used as the benchmark model
2. Single decision tree
3. Ensemble model combining the above 2 classifiers to form a logistic model tree: LMT

[1] A HELOC is a line of credit typically offered by a bank as a percentage of home equity (the difference between the current market value of a home and its purchase price)

4. Ensemble model based on gradient tree boosting: XGBoost

The above models were selected both for the interpretability they offer and the promising performance they have displayed in credit scoring exercises.

Once the explainability criteria of each model has been established, this will guide the subsequent model optimisation process to obtain the best predictive performance. The best performing models from each of the 4 categories will be assessed in terms of explainability and predictive performance to pick the final recommendation.

5. Benchmark Model

Given the dual goal of accuracy and explainability, the logistic regression model will be used as benchmark as its output is a linear function of the input variables. The sign of each feature's coefficient can be analysed relative to the monotonicity constraint provided in the data dictionary to ensure that the expected patterns are respected.

6. Evaluation Metrics

In the context of the HELOC dataset, both false negatives and false positives lead to unfavourable outcomes. False negative results in the refusal of credit to a good customer, with the subsequent loss of interest income. This will be monitored via the *sensitivity* measure.

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

False positive results in incorrectly classifying a bad RiskPerformance as good and thereby granting credit to a bad customer on favourable terms. This can result in loss of the loan principal as well as any additional costs arising from the loan default. This will be monitored via the *specificity* measure.

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

Both sensitivity and specificity will be combined in a single score (G-mean). The larger the G-mean, the superior is the model.

$$G = \sqrt{sensitivity \times specificity}$$

As far as explainability is concerned, there is no quantitative metric to measure its effectiveness. Instead, the expectation is that both the global and local explanations can be read and fully comprehended by a human (in this case a FICO data scientist) in under 10 mins.

7. Project Design

The project will be structured as follows:

a. **Exploratory Data Analysis:** The individual features will be analysed via visualisations to discern any correlations and patterns with the target variable. Data imputation will be performed to fix any gaps. New features will potentially be derived off the existing features if they are deemed to be relevant for predicting RiskPerformance. The importance of the features will be assessed via PCA.
b. **Model Explainability Approach:** For each of the chosen models, a suitable framework will be put together to generate local and global explanations. This will be a mixture of existing scikit functions as well as more sophisticated approaches (e.g SHAP).
c. **Model Fitting and Optimisation:** Each of the 4 models will be fitted to the dataset using cross-fold validation and the model hyperparameters will be tuned to maximise performance without overcomplicating the model.
d. **Evaluation of Optimised Models:** The results, including the explanations, from the optimised models will be analysed in the context of the credit scoring exercise. The recommended model will be chosen based on its prediction performance, the simplicity and transparency of its explanations and the plausibility of its interpretations.

8. References

- FICO Explainable Machine Learning Challenge. https://community.fico.com/community/xml/pages/overview
- David Gunning, DARPA. Explainable Artificial Intelligence (XAI). 2017.
- Niels Landwehr, Mark Hall, Eibe Frank. Logistic Model Trees. 2004.
- Tianqi Chen, Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. 2016.
- Finale Doshi-Velez, Been Kim. Towards a Rigorous Science of Interpretable Machine Learning. 2017.
- FICO. Understanding FICO Scores. https://www.myfico.com/Downloads/Files/myFICO_UYFS_Booklet.pdf
- Shunpo Chang, Simon Dae-Oong Kim, Genki Kondo. Predicting Default Risk of Lending Club Loans. 2015.
- Junjie Liang. Predicting borrower's chance of defaulting on credit loans. 2011.
- Marie-Laure Charpignon, Enguerrand Horel, Flora Tixier. Prediction of Consumer Credit Risk. 2014.
- Yang Liu. The evaluation of classification models for credit scoring. 2002.
- Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, Finale Doshi-Velez. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. 2018.
- Marco Ribeiro, Sameer Singh, Carlos Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. 2016.
- Christoph Molnar. Interpretable Machine Learning A Guide for making black box models interpretable. 2018.
- Scott Lundberg, Su-In Lee. A Unified Approach to Interpreting Model Predictions. 2017.