Loan Prediction

Machine Learning Engineer Nanodegree

Capstone Proposal

Elvin Rey Magsino

August 27, 2017

Proposal

Domain Background

The insurance domain is one of the largest consumers of data science analytics. This industry is founded on forecasting and estimating the value/impact of future events. Insurance has been using established predictive modeling practices for some time now, with predictive modeling being used especially in claims loss prediction and pricing. The opportunity to use machine learning techniques, even across new insurance operations, has never been greater as new data sources and big data are more readily available. Machine learning allows insurance companies to yield higher predictive accuracy since it can fit data using more flexible and complex models. ML takes advantage of heavy data analytics and is capable of computing datasets that are seemingly unrelated in a structured, semi-structured, or non-structured environment. Insurers can use ML to better understand the core values of insurance such as risk, claims, and customer experience. Critical areas where machine learning can greatly benefit insurance includes estimating losses, monitoring and detecting fraud, claims processing, and pricing risk.

Gary Reader, KPMG's Global Head of Insurance, one stated,

> *For the insurance sector, we see machine learning as a fundamental game-changer since most insurance companies today focus on three main objectives: improving compliance, improving cost structures and improving competitiveness. Machine learning can form at least part of the answer to all three.*[1]

In a business sense, this is why ML drives so much value:

- *Machine learning delivers more accurate predictions than traditional analysis or human judgment.*
- *Modern techniques make these predictions easy to understand and transparent.*
- *With better predictions, managers make smarter decisions.*
- *Smarter decisions produce more revenue, lower costs, and a better bottom line.*[1]

For this project, we'll be analyzing Dream Housing Finance (DHF), an insurance company which deals specifically with home loans. To calculate loan eligibility, Dream Housing Finance asks potential customers a few questions ranging from income level to loan amount. This insurance company has a presence across all types of property areas – urban, semi urban, and rural. Machine Learning techniques will be implemented to determine the customer's loan eligibility.

Problem Statement

When a customer provides their personal details (including gender, marital status, education, number of dependents, income, education level, loan amount, loan term, and credit history), DHF will automatically generate the customer's loan eligibility. Since the output of the algorithm is either a yes or no, this poses a binary classification problem. DHF will need to implement machine learning algorithms to accurately predict whether a certain customer will be approved for a home loan.

Datasets and Inputs

Data was extracted from the loan prediction dataset available through Analytics Vidhya[3]. This project will consider a dataset consisting of 981 rows and 14 categories (description, unique load ID, male/female, applicant married (Y/N), number of dependents, applicant education (graduate/under graduate), self-employed, applicant income, coapplicant income, loan amount in thousands, term of loans in months, credit history meets guidelines, urban/semiurban/rural, loan approved (Y/N). Each row represents one customer whose target variable, loan approved (Y/N), is what the algorithm will be predicting. The target variable can take two values, either yes or no; therefore, this will be a binary classification problem.

Since the distribution of the target variable is unbalanced (of the 615 requests, 422 were granted for a total of 68.73% success), we would need a strategy to deal with this unbalanced data. The evaluation metric will be an accuracy score – the percentage of the loan approval that the algorithm correctly predicts. However, due to unbalanced data, we're going to look at different performance measures that can give more insight to the accuracy of the model than traditional classification accuracy – this includes precision, recall, and the F-Score (a weighted average of precision and recall).

Solution Statement

We will use a generic classification function, which takes a model as input and determines the accuracy and cross-validation scores. We will also experiment with different models to determine which ones offer better results. (It should be noted however, that using a more sophisticated model does not necessarily mean it will yield better results.)

Looking at and working with the different categories in the dataset, we'll be able to get a deeper understanding of which variables place a higher emphasis on whether or not a loan will get approved. During the solution phase, we are going to determine if an interpretable algorithm (such as a decision tree or logistic regression) would be more appropriate to implement to this dataset than a more complex, higher accuracy one (boosting or neural networks). In general, decision trees often perform well on imbalanced data – so I initially suspect we'll get higher accuracy with that model in paticular[2].

Benchmark Model

For this project, I will be utilizing at least three models. One model I will implement is logistic regression. Since this is the most basic type of regression and commonly used in predictive analysis, its results will let us examine a few things. Does a set of predictor variables do a good job in predicting an outcome variable? Is the model using the predictors accounting for the variability in the changes in the dependent variable? Which variables in particular are significant predictors of the dependent variable? And in what way do they impact the dependent variable? However, since logistic regression is sensitive to outliers, this could significantly swing the regression results. Additionally, logistic regressions tend to overfit the data and can begin to model the random noise in the data, rather than just the relationship between the variables.

After logistic regression, we'll apply a better modeling technique – decision trees. Decision trees are known to provide higher accuracy than regression models. Unlike logistic models, decision trees can map non-linear relationships quite well.

We'll be using random forests since it's capable of both regression and classification tasks – it also undertakes dimensional reduction methods, treats missing values, outlier values, and other steps of data exploration, and does a fairly good job. It is a type of ensemble model where a group of weak models combine to perform a powerful model.[6] I suspect random forest to yield the best results due to being able to work with all the features in the dataset and returning a feature importance matrix which we can use to select important features. Additionally, random forest has methods for balancing errors in data sets where classes are imbalanced.

Using these methods will allow us to compare their accuracy and cross validation scores and select the most appropriate model to predict loan approval.

Evaluation Metrics

We'll be looking at the accuracy of the target Boolean variable, Loan_Status, to quantify the performance of both the benchmark model and the solution model. Since the distribution of the target variable is unbalanced (of the 615 requests, 422 were granted for a total of 68.73% success), we would need a strategy to deal with this unbalanced data. The evaluation metric will be an accuracy score – the percentage of the loan approval that the algorithm correctly predicts. However, due to unbalanced data, we're going to look at different performance measures that can give more insight to the accuracy of the model than traditional classification accuracy – this includes precision, recall, and the F1-Score (a weighted average of precision and recall).

We'll be interested in the F1 score as the measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results, and r is the number of correct positive results divided by the number of positive results that should have been returned. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst at 0.[4]

Additionally, another evaluation metric we can possibly use is ROC (receiver operating characteristic) score – which calculates the area under the ROC curve. In an ROC curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points of a parameter. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups (diseased/normal).[5]

(The equation used for calculating the F1 score)

$$F_1 = 2 \cdot \frac{1}{\frac{1}{\mathrm{recall}} + \frac{1}{\mathrm{precision}}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

Project Design

- Language: Python 3.6.1
- Libraries: NumPy, Pandas, Matplotlib, Scikit-learn
- Outline:
    - Import libraries and dataset, visualize, and explore data
    - Study distribution of various variables (look at extreme values, missing, erroneous, or out-of-place data)
    - Baseline calculations – Calculate probability of loan from various classes (credit history, education, dependents, etc).
    - Clean up data – solve missing values in the dataset, solve unpractical values in the dataset (replacing missing values with the mean for example, or assigning a log function to extreme values to nullify effect)
    - Build predictive models – encode labels (for nonnumerical data)
        - Apply logistic regression and note its accuracy and various scores
        - Apply decision tree and note its accuracy and various scores
        - Apply random forest and note its accuracy and various scores
    - Conclude which model yields best results

Citations

[1]Sengupta, Satadru. “The Power of Machine Learning in Insurance.” Cloudera VISION, Cloudera, 11 May 2017, vision.cloudera.com/the-power-of-machine-learning-in-insurance/.

[2]Brownlee, Jason. “Machine Learning Algorithm Recipes in Scikit-Learn.” Machine Learning Mastery, Machine Learning Mastery, 21 Sept. 2016, machinelearningmastery.com/get-your-hands-dirty-with-scikit-learn-now/.

[3]“Practice Problem: Loan Prediction III | Knowledge and Learning.” DataHack : Biggest Data Hackathon Platform for Data Scientists, Analytics Vidhya, 25 May 2016, datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/.

[4] “F1 Score.” Wikipedia, Wikimedia Foundation, 28 Aug. 2017, en.wikipedia.org/wiki/F1_score.

[5] Schoonjans, Frank. “ROC Curve Analysis with MedCalc.” MedCalc, MedCalc Software, 9 May 2017, www.medcalc.org/manual/roc-curves.php.

[6]Team, Analytics Vidhya Content, et al. “A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python).” *Analytics Vidhya*, Analytics Vidhya, 1 May 2017, www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/.