**All-State Loss Prediction**

**Akash Deo(apd160330) | Mohammad Atif Hussain(mxx160130)**

**University of Texas at Dallas**

1. **Introduction**

All State was founded on April 17,1931 as an auto insurance company that offered insurance through the mail. In 1941, only 25% of drivers had auto insurance which later lead to all other states requiring drivers to have auto insurance. In the insurance business, many insurance companies want to determine the amount of money they will lose for a claim given a certain customer. This is important for an insurance company to be able to predict the amount they will lose because if they cannot determine the amount then there is a chance that said company may go out of business because of too many losses. Our basic approach is to find a pattern in the Categorical and Continuous data that is given to us to determine how much the company will lose given a set of attributes for a customer.

1. **Problem Definition and Algorithm**
   1. **Task Definition**

The problem we are addressing is given a set categorical data and continuous data we have to accurately predict the loss for an insurance company. We have to train a machine learning algorithm given a set of training data and to test how accurate the algorithm is given a set of testing data. This is important because it can predict the amount lost from an insurance claim given a set of attributes for a customer.

* 1. **Algorithm Definition**

Two Algorithms were used, a non-linear Support Vector Machine and Random Forest which is an ensemble method for regression. In both models we are trying to find the best parameters that give us a reasonable R-Squared values to see how overfit or underfit our model is, where R-Squared uses the formula where is the average of the target value, o is the output of the predictor, and t is the target. The preprocessing for both algorithms involved removing any data that was unnecessary. We achieved this by finding all correlations between the attributes and the class as well as any correlations between the attributes themselves. Once this was plotted we removed any attributes that were heavily related to each other and any attributes that had a correlation below the threshold 0.01 with respect to the loss value.

* + 1. **SVM(Non-linear)**

A Support Vector Machine is a machine learning algorithm that tries to find the best hyper-plane that represents the training data with the best balance between variance and bias to obtain the best possible R-Squared value. In the SVM algorithm, we used the Polynomial kernel which is of the form, , where γ is a user defined as , c is a parameter needed for polynomial kernels, u’ and v are data instances, and d is the degree of the polynomial. We then split the data into 10 sets of 20,000 instances where the last set only contained about 9,000 instances. We then built 10 different SVM models but using the sets as training datasets. Once the SVM models are trained we take all the training data and obtain a prediction of the loss value for all instance in each training model. Then we find the average of the predicted value for each data instance from all 10 different models and take that as the final loss prediction. After we find all the loss predictions we find the R-Squared value to see how overfit or underfit the model is.

* + 1. **Random Forest**

A Random Forest is a machine learning algorithm that involves bootstrapping and randomly selecting an attribute for each tree model. Specifically, the algorithm creates a bootstrap sample of some size N, grows a tree using a random selection of attributes from the sample. After the tree is grown depending on if the problem is a classification or regression problem it will choose majority or average, respectively. Similarly, we split the data into 10 sets of 20,000 instances where the last set only contained about 9,000 instances. We again created 10 different models based on the training data that was provided to us. Once the Random Forest models were trained we took the training data and used it to test each model that was created and predicted the loss for instance in each model. Then we took the average of all the predicted losses and used it as our final prediction for each instance. After we found all the predicted losses we found the R-Squared value to see how overfit or underfit the model is to the training data.

1. **Experimental Evaluation**
   1. **Methodology**

The specific hypothesis we are trying to test is if the change in categorical and continuous data changes the amount of loss that results. The independent variable that are present in the training and test is 116 categorical data types and 14 continuous data types. There was only one dependent variable which was the losses from each claim instance. The problem can be considered a regression problem because loss amount is a continuous data type. Regression analysis helps one see that the value of the dependent variable can change when any of the independent variables vary. The data that we were given is interesting because the attributes for the training and test data are not given a descriptive label. Also in the test data the loss is not given to us so we must predict the loss given a set of attributes. The performance data we obtained from the test was the R-Squared Error which indicates the proportion of the variance in the dependent variable that is predictable from the independent variable.

Before Preprocessing:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID | A1 | A2 | A3 | A4 | Loss |
| 1 | A | A | C | 0.16 | 0 |
| 2 | B | B | A | 0.26 | 12000 |
| 3 | A | B | A | 0.25 | 30000 |
| 4 | B | C | AD | 0.67 | 20000 |

After Preprocessing:

|  |  |  |
| --- | --- | --- |
| A1 | A3 | Loss |
| A | C | 0 |
| B | A | 12000 |
| A | A | 30000 |
| B | AD | 20000 |

After Processing:

R-Squared: 0.579

* 1. **Results**
     1. **SVM(Non-Linear)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Kernel | Degree | Coef0 | Gamma | Epsilon | Cost | R-Squared |
| Polynomial | 2 | 0.5 | 1/94 | 0.2 | Cost | 0.5667128 |
| Polynomial | 2 | 05 | 0.005 | 0.2 | 10 | 0.532164 |
| Polynomial | 3 | 0 | 1/94 | 1 | 10 | 0.4230988 |
| Polynomial | 3 | 0.1 | 1/94 | 1.5 | 10 | 0.1085126 |

* + 1. **Random Forest**

|  |  |  |  |
| --- | --- | --- | --- |
| Mtry | Ntree | NodeSize | R-Squared |
| 15 | 700 | 4 | 0.5790 |
| 20 | 500 | 4 | 0.5786 |
| 10 | 500 | 2 | 0.577 |
| 4 | 1500 | 3 | 0.52 |
| 20 | 500 | 3 | 0.58 |

* 1. **Discussion**

We believed that the Random Forest algorithm would perform better than the SVM algorithm simply because it is an ensemble method which we learned is the best Machine Learning algorithm to use. From the results that we got our hypothesis is correct and it shows that SVM may be a powerful algorithm but it is not as powerful as an ensemble method. The weakness of the SVM is the fact that with the variables we used the parameters formed a 94-dimensional system which would have a hard time trying to achieve higher performance than a Random Forest decision tree. The ensemble method we used, Random Forest, uses decision trees which has an easier time finding higher performance because it mainly many perceptrons.

1. **Related Work**

Predicting Future Paid Losses:

The problem that the authors are trying to solve is to estimate the amount of losses that a state must pay in insurance losses given historical data. Their method of predicting the losses are if the datasets have a strong correlation then the predicted value will be close to the immediate past value. However, if they are not strongly correlated then the predicted value will be the mean of all past values. The authors’ work differs from ours in that their predictions were used for all insurance companies in a state rather than an individual insurance company. Our problem is better because their work consists of multiple insurance companies that each may have different policies while ours consists of a single company with a single set of policies. Also, the authors used basic math equations to predict their losses while we used Machine Learning Algorithms to predict our losses.

1. **Future Work**

Our datasets have many different instances in it which also include some outliers. One example is the instance that has a loss of approximately 120,000. This outlier increases our error in our models which can have two different was of being dealt with. The first way is to eliminate the outlier so that our error decreases. However, we do not want to throw away data so this would not be desirable. The second way would be to replot the point by finding a similar instance in the dataset and recasting the loss of the outlier to be the same as the loss of the similar point. Another possible improvement would be to use a different machine learning algorithm. One possible algorithm would be the Gradient Boosting algorithm which another UTD student used in a different insurance loss prediction problem.

1. **Contribution of team-members.**

Data Preprocessing techniques were discussed among all the members and was implemented by Atif, SVM was implemented by Phalguna and Matthew and RandomForest was implemented by Kajol and Akash.

1. **Conclusion**

In this paper, we applied the machine learning algorithms we learned in our class to determine which Machine Learning Algorithm would be the best for a real-life problem. Even though our results were around 50-60 percent accurate for R-squared values these results were very good. Looking at the two algorithms we used to predict the losses we found that Random Forest is the better than SVM in predicting values.

In the future, if we are to come back to this problem again we will have a better starting point than when we started this for the first time. The work presented in this paper will help form a starting point on which Machine Learning Algorithm will perform with better accuracy given the ones used in our paper.

**Bibliography**

Liu, Haiying et al. “Predicting Future Paid Losses.” *Worcester Polytechnic Institute*, Hanover Insurance Group, web.wpi.edu/Pubs/E-project/Available/E-project-043009-161344/unrestricted/Predicting\_Future\_Paid\_Losses.pdf.

“Loss Function.” *Wikipedia*, Wikimedia Foundation, en.wikipedia.org/wiki/Loss\_function.

“References.” *Support Vector Machine Regression*, kernelsvm.tripod.com/.

“Support Vector Regression with R - SVM Tutorial.” *SVM Tutorial*, 2016, www.svm-tutorial.com/2014/10/support-vector-regression-r/.

“SVM in R for Regression.” *Stack Overflow*, stackoverflow.com/questions/8014730/svm-in-r-for-regression.

Institute, By DnI. “DnI Institute.” *DnI Institute*, May 2015, dni-institute.in/blogs/random-forest-using-r-step-by-step-tutorial/.

Breiman, Leo et al. “Package 'RandomForest'.” *Cran-R*, Cran-R, 7 Oct. 2015, cran.r-project.org/web/packages/randomForest/randomForest.pdf.

Liaw, Andy, and Matthew Weiner. “Classification and Regression by RandomForest.” *Biostatistics UNC*, UNC, Dec. 2002, www.bios.unc.edu/~dzeng/BIOS740/randomforest.pdf.

Quantsignals. “Learning Kernels SVM.” *R-Bloggers*, 25 Sept. 2012, www.r-bloggers.com/learning-kernels-svm/.

Sastry, P. S. “Support Vector Regression and ?-Insensitive Loss Function, Examples of SVM Learning | Pattern Recognition.” *Support Vector Regression and ?-Insensitive Loss Function, Examples of SVM Learning | Pattern Recognition*, www.myopencourses.com/video/support-vector-regression-and-insensitive-loss-function-examples-of-svm-learning.

Tiwari, Rajdev, and Manu Pratap Singh. “Correlation-Based Attribute Selection Using Genetic Algorithm.” *IJCA Online*, International Journal of Computer Applications, Aug. 2010, www.ijcaonline.org/volume4/number8/pxc3871182.pdf.

Tiwari, Rajdev, and Manu Pratap Singh. “Correlation-Based Attribute Selection Using Genetic Algorithm.” *International Journal of Computer Applications*, vol. 4, no. 8, Aug. 2010, pp. 28–34.