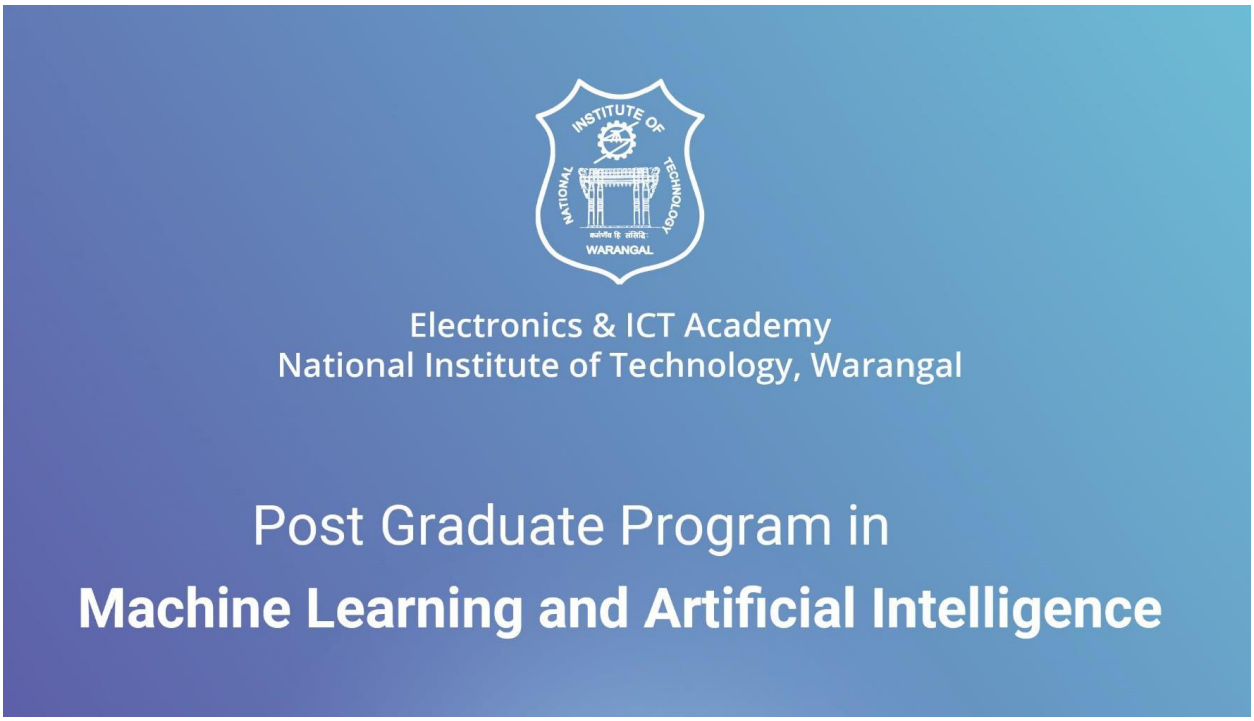

Project Synopsis

Submitted by

Jiss Peter

Technical Architect – Tata America International Ltd (TCS Ltd.)

Jiss07@gmail.com

+1(754)-248-3941

Aim of the Project

The aim of the project is to build a Machine Learning Model to predict whether an owner will initiate an auto insurance claim in the next year

Background

The auto insurance industry is witnessing a paradigm shift. Since auto insurance company consists of homogenous good thereby making it difficult to differentiate product A from product B, also companies are fighting a price war (for insurance price). On top of that, the distribution channel is shifting more from traditional insurance brokers to online purchases, which means that the ability for companies to interact through human touchpoints is limited, and customers should be quoted at a reasonable price. A good price quote is one that makes the customer purchase the policy and helps the company to increase the profits.

Problem statement

Nothing ruins the thrill of buying a brand new car more quickly than seeing your new insurance bill.

The sting's even more painful when you know you're a good driver. It doesn't seem fair that you have to pay so much if you've been cautious on the road for years.
The problem statement for the project is to create a suitable model which will predict the probability that an auto insurance policy holder files a claim. You will be given a train.csv, a dataset with 600k training data and 57 features/data.

In the train and test data, features that belong to similar groupings are tagged as such in the feature names (e.g., ind, reg, car, calc)

Solution / Architecture used

Various techniques used for reaching out to the final solution is summarized as given below

1. Data Pre processing and Feature engineering.
2. Categorization of features and finding out correlation coeff to do the feature selection and also data cleansing by removing or replacing missing values and NaN values
3. Encoding and sampling techniques has been applied to define a better data set for feeding in to algorithms
4. Data classification using below given algorithms

Random Forest Classifier

It is an ensemble tree-based learning algorithm. The Random Forest Classifier is a set of decision trees from randomly selected subset of training set. It aggregates the votes from different decision trees to decide the final class of the test object.

XGBoostClassifier

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now.

Logistic Regression

Logistic Regression is a 'Statistical Learning' technique categorized in 'Supervised' Machine Learning (ML) methods dedicated to 'Classification' tasks. It has gained a tremendous reputation for last two decades especially in financial sector due to its prominent ability of detecting defaulters.

MLP Classifier

MLP Classifier stands for Multi-layer Perceptron classifier which in the name itself connects to a Neural Network. Unlike other classification algorithms such as Support Vectors or Naive Bayes Classifier, MLPClassifier relies on an underlying Neural Network to perform the task of classification.

SVM Classifier

In machine learning, support-vector machines (SVMs, also support-vector networks[1]) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier

Exploratory Data Analysis Questionnaire and Answers:

1. **Write at least 3 important inferences from the data above(train.csv)**

- The given dataset contains 595212 rows of data with 59 columns. The target variable is in the form of 0 and 1 and all other variables are numeric and there by the dataset is suitable for Exploratory data analysis
- The target variable contains only 0 and 1.But this variables are not uniformly distributed and there by balancing is required on this target variable
- The dataset contains missing values which needs to be addressed and also categorical features and binary features and numerical features are available in the dataset.

2. **Is the data balanced? Meaning are targets 0 and 1 in the right proportion?**

- No. The Data is not balanced for the train.csv file given. While doing visualization of the target variable proportion and distribution as a pie chart we have observed it clearly that the value 0 is having major portion in the target variable as compared to the other value available as 1 .Inorder to overcome this issue we have applied sampling technique and has balanced the target variable

3. **How many categorical features are there?**

   There are mainly 14 Categorical features available in the dataset and is as given below

- ['ps_ind_02_cat', 'ps_ind_04_cat', 'ps_ind_05_cat', 'ps_car_01_cat', 'ps_car_02_cat', 'ps_car_03_cat', 'ps_car_04_cat', 'ps_car_05_cat', 'ps_car_06_cat', 'ps_car_07_cat', 'ps_car_08_cat', 'ps_car_09_cat', 'ps_car_10_cat', 'ps_car_11_cat']

4. **How many binary features are there?**

- There are mainly 17 Binary features available in the dataset and is as given below
- ['ps_ind_06_bin', 'ps_ind_07_bin', 'ps_ind_08_bin', 'ps_ind_09_bin', 'ps_ind_10_bin', 'ps_ind_11_bin', 'ps_ind_12_bin', 'ps_ind_13_bin', 'ps_ind_16_bin', 'ps_ind_17_bin', 'ps_ind_18_bin', 'ps_calc_15_bin', 'ps_calc_16_bin', 'ps_calc_17_bin', 'ps_calc_18_bin', 'ps_calc_19_bin', 'ps_calc_20_bin']

5. **Write inferences from data on interval variables.**

The numerical features consist of interval and ordinal variables. We have tried to find out how many missing values there are for each feature type and has tried to Eliminate the features where more than one half of the values are missing. What we have observed is that the inetrval variables have less amount of missing data and the same has been substituted with applicable mean.

6. **Write inferences from data on ordinal variables.**

Even though there are multiple ordinal values available in the given dataset the margin of missing values is very less. We have observed only one ordinal variable with more than 50 % of missing value as ps_car_03_cat.

7. **Write inferences from data on binary variables.**

For the binary variables no columns were returned where more than half of there values missing. we also made sure that no values at all are missing for the binary features.

8. **Check if the target data is proportionate or not.Hint: Below than 30% for binary data is sign of imbalance**

The target variable is not proportionate. Initially when we tried to plot a pie chart it was evident that only 3.6 of the target variable is havinga value of 1 and rest all values where 0s.

9. **What should be the preferred way in this case to balance the data?**

For doing balancing first we have figured out the indices of the unbalanced data per target indices and has randomly selected records with target=0 to get at the desired a priori, then calculated the undersampling rate and resulting number of records with target=0.Constructed a list with remaining indices and returned undersample data frame back to train dataset.The final result was as follows

* Rate to undersample records with target=0 :0.34043569687437886
* Number of records with target=0 after undersampling: 195246

10. **How many training records are there after achieving a balance of 12%?**

After balancing and data pre processing the resulted dataframe [df_rebalanced] was having the shape as (579628, 47) ahead of the original dataframe [df] of (595212, 59). That means , some rows where added , but total number of columns reduced to 47 from 59.

11. **Which are the top two features in terms of missing values?**

The top two columns with missing values are as given below * col1 : ps_car_03_cat missing count : 4,11,231 * col2 : ps_car_05_cat missing count : 2,66,551

12. **In total, how many features have missing values?**

In total 12 columns have missing values and they are as given below

Column----------> Missing count----->Missing ratio

ps_ind_02_cat ---> 216 ---> 0.000

ps_ind_04_cat ---> 83 ---> 0.000

ps_ind_05_cat ---> 5809 ---> 0.010

ps_car_01_cat --->107 ---> 0.000

ps_car_03_cat --->411231 --->0.691

ps_car_05_cat --->266551 --->0.448

ps_car_07_cat --->11489 --->0.019

ps_car_09_cat --->569 ---> 0.001

ps_reg_03 --------> 107772 ---> 0.181

ps_car_11 --->5 ---> 0.000

ps_car_12 --->1 ---> 0.000

ps_car_14 --->42620 --->0.072

13. **What steps should be taken to handle the missing data?**

We have substituted the missing values with the applicable column mean. This have limited their impact on the results.

14. **Which interval variables have strong correlation?**

ps_reg02 and ps_reg03 have highest correlation among interval variables

15. **What's the level of correlation among ordinal features?**

The correlations are very small, so not worthy of consideration.

16. **Implement Hot Encoding for categorical features**

We have removed most of the categoricalfeartures after feature engineering and tehre by the resulting feature list or categorical columns doesn't require one hot enncoding.

17. **In nominal and interval features, which features are suitable for StandardScaler?**

Scaling features tends to lead to a performance improvement with classification problems, and hence we taken both interval and nominal features in to consideration.

18. **Summarize the learnings of ED**

The given dataset contains 595212 rows of data with 59 columns. But the dataset had many missing values and Nan values as well as the target variable was not uniformly distributed. Majority of the missing values on ordinal and interval variable has been replaced with corresponding mean and for categorical features & binary features with missing values low person coeff has been removed. Also done balancing on the target variable as well as applied sampling and unsampling techniques for the given dataset to create a finalized dataframe and values has ben scaled using MinMaxScaler for ingesting the same in to various classification algorithms. The final data has been split using train_test_split function and has feeded to

classification algorithm and prediction has been made and also the accuracy has been calculated and printed.

Modelling Questionnaire and Answers:

1. **The Simple Logistic Regression Model seems to have high accuracy. Is that what we need at all? What is the problem with this model?**

   I have tried to fit the given data with Logistic regression model. But the accuracy was very low as compared to Random Forest Classifier/Accuracy was found to be nearly 0.59 or 0.62 for both of the target variables . Even XGBoost and Linear SVM where also having low accuracy for prediction.

2. **Why do you think f1-score is 0.0?**

   An F1-score means a statistical measure of the accuracy of a test or an individual. It is composed of two primary attributes, viz. precision and recall, both calculated as percentages and combined as harmonic mean to assign a single number, easy for comprehension.F1 score 0.0 means that the both the testing and training accuracy was poor and it indicates under fitting.

3. **What is the precision and recall score for the model?**

   I have tried Random Forest, XGBoost, Logistic regression, Linear SVM and MLP classifier for the given dataset. Corresponding precision and recall and accuracy is as given below
   Algorithm: Random Forest --> Precision : 1, Accuracy : 1, Recall : 1
   Algorithm: XGBoost --> Precision : 0.59 , Accuracy : 0.61, Recall : 0.62
   Algorithm: Logistic Regression --> Precision : 0.59, Accuracy : 0.58, Recall : 0.58
   Algorithm: LinearSVC --> Precision : 0.46, Accuracy : 0.46, Recall : 0.46
   Algorithm: MLPClassifier --> Precision : 0.79, Accuracy : 0.78, Recall : 0.79

4. **What is the most important inference you can draw from the result?**

   I personally felt like the Random Forest algorithm suits best for the given dataset based on the precision and accuracy rates. XGBoost and Linear SVM as well plays an important role , but accuracy is low

5. **What is the accuracy score and f1-score for the improved Logistic Regression model?**

   Algorithm: Logistic Regression --> Precision : 0.59, Accuracy : 0.58, Recall : 0.58

6. **Why do you think f1-score has improved?**

   F-measure or balanced F-score (F1 score) is the harmonic mean of precision and recall.Logistic regression model uses cumulative gain approach , which can improve the f1 score by increasing the precision and recall rate.

7. **For model LinearSVC play with parameters – dual, max_iter and see if there is any improvement**

   Tried the same , but no such improvements has been observed

8. **SVC with Imbalance Check & Feature Optimization & only 100K Records → is there improvement in scores?**

   I havn't tried this

9. **XGBoost is one the better classifiers -- but still f1-score is very low. What could be the reason?**

   I agree to the fact that the XGBoost is a better classifier. But for the given dataset the precision and recall rate was very low.As F1 score is a direct measure of these two parameters , the score has went down.

10. **What is the increase in number of features after one-hot encoding of the data?**

    After the encoding the number of features got almost doubled

11. **Is there any improvement in scores after encoding?**

    Yes. Encoding has improved teh classification accuracy

12. **If not missing a positive sample is the priority which model is best so far?**

    RandomForest Algorithm and XGBoostalgoritms seems to be better for the classification in this case.

13. **If not marking negative sample as positive is top priority, which model is best so far?**

    RandomForest Algorithm have the better for the classification in this case.

14. **Do you think using AdaBoost can give any significant improvement over XGBoost?**

    Yes. Adaboost classifier have better recall and accuracy than XGBoost for the given dataset

15. **MLPClassifier is the neural network we are trying. But how to choose the right no. of layers and size?**

    Default layers is 100 for MLP classifier. We need to do trial and error method to find out the best layers and size.

16. **At what layer size we get the best f1-score?**

    The default number of hidden layer is 100.I have tried the MLP classifier with 50 , 100 and 150 layers and found out that the default layer of 100 gives best accuracy and f1 score.

Real World applications / business use cases

1. **Conquering Market Share:** Capture market share by lowering the prices of the premium for the customers, who are least likely to claim.
2. **Risk Management:** Charge the right premium from the customer, who is likely to claim insurance in the coming year
3. **Smooth Processing:** Reduce the complexity of pricing models. Most of the transactions are happening online with larger customer attributes (thanks to the internet and social media). Harness the power of huge data to build complex ML models
4. **Increased Profits:** As per industry estimate 1% reduction in the claim can boost profit by 10%. So, through the ML model, we can identify and deny the insurance to the driver who will make a claim. Thus, ensuring reduced claim outgo and increased profit.

Summary

In this project, we have been challenged to build a model that predicts the probability that a driver will initiate an auto insurance claim in the next year. A more accurate prediction will allow them to further tailor their prices, and hopefully make auto insurance coverage more accessible to more drivers. The dataset was really challenging as it had multiple missing and unbalanced data.Exploratory data analysis or EDA has helped us to do a proper processing of various features and to understand the data in depth and also we could be able to visualize each of teh individual features in depth.We got an opportunity to apply various algorithms such as RandomForest,XGBoost,SVM,MLP,Logistic Regression etc and could be able to compare the results of each of the algorithms to reach out to a conclusion on the better algorithm to be picked up.Overall the project has given an opportunity to apply multiple deep learning skills in a single project.

Conclusion

We have done EDA, Feature engineering, Feature selection, visualization of teh features and also applied various algorithms and could be able to predict the possiblity of a given driver to be safe or not. Multiple models have been created and teh best one will be picked up and will be used for further implementation.

References

[1] https://www.kaggle.com/c/porto-seguro-safe-driver-prediction
[2] Xianzhi Liu, Qingquan Song - Safe Driver Prediction Base on Different Machine Learning Models
[3] Feiyang Pan,Xiang Ao - Institute of Computing Technology -A Simple and Empirically Strong Method for Reliable Probabilistic Predictions
[4] Muhammad Arief Fauzan, Hendri Murfi -Department of Mathematics, Universitas Indonesia - The Accuracy of XGBoost for Insurance Claim Prediction
[5] Matthew Millican & Laura Zhang - CS 229 Final Report: Predicting Insurance Claims in Brazil