**CSC 529**

**Porto Seguro’s Safe Driver Prediction**

**Detecting Auto Insurance Claims**

**Team members:**

**Kai Chung Ying - online student**

**Tyler Jewell - online student**

**Abstract:**

**Insurance companies formulate insurance policy based on factors that are related to driver’s drivability or vehicles’ specification etc. Inaccurate accounting those elements influence customers’ premium or intensify the operation cost of insurance companies. One of insurance company from Brazil, Porto Seguro, offers an opportunity for us to build a model to predict policyholders would file claims in the following year. This problem will be** accomplished **by a number of Data Mining and Machine Learning techniques in the project.**

**Introduction and motivation:**

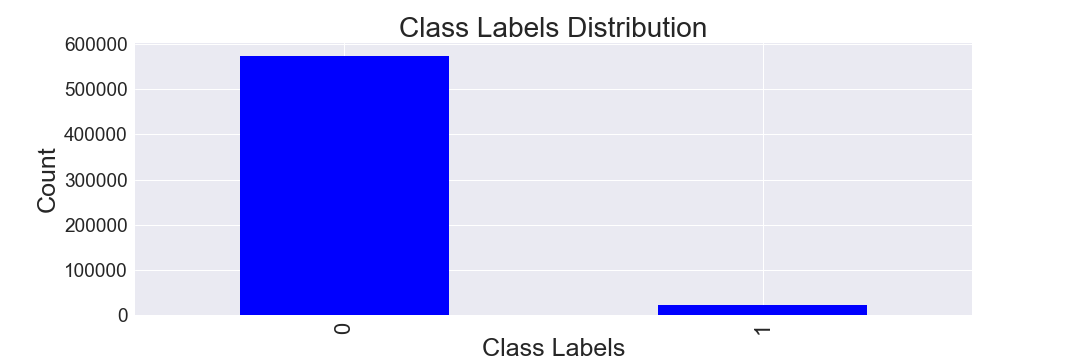
**The objective of this project is to predict whether an auto insurance policy holder will file a claim next year by using the Porto Seguro’s Safe Driver Prediction dataset from Kaggle. All insurance companies are very interested in knowing what factors cause their insurance holders to file a claim. This can be used to adjust premiums based off of risk and also minimize the most important factors that cause a person to make a claim.**

* With the given dataset, would our technique / models be able to predict correctly whether the policy holders would initiate an auto claim in the next year which refers to Class label equals “1”?
* •Which classifiers/ models be able to efficiently determine the highest accuracy of this classification experiment?

**Data Analysis**

**The dataset for this Binary classification study is downloaded from the Kaggle.com** <https://www.kaggle.com/c/porto-seguro-safe-driver-prediction/data>**. The dataset is provided through Kaggle.com and prepared by Porto Seguro which is one of Brazil’s largest auto and homeowner insurance companies. Each instance of the data corresponds to the information of one policy holder. The Target attribute which we are predicting contains TWO class labels: “1” means that claim was filed and “0” means that No claim was filed. The complete dataset is composed of** 595212 instances and 57 attributes with no duplicated instances. Upon our cross-checking, missing values of ordinal and categorical values are replaced by Mode value of the corresponding variable and missing values of numeric variables are replaced by the mean of the corresponding variable. The dataset is highly biased towards the “0” class which means that particular person did not file a claim as shown on the figure 1 below. Unfortunately, there is no detail description provided for each attribute except the data type of variables which is listed table 1:

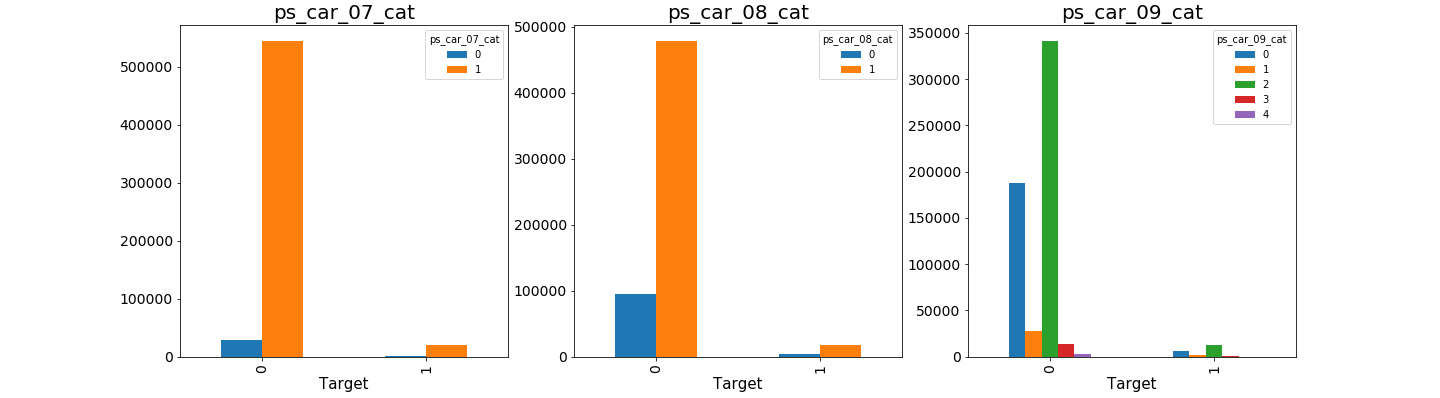
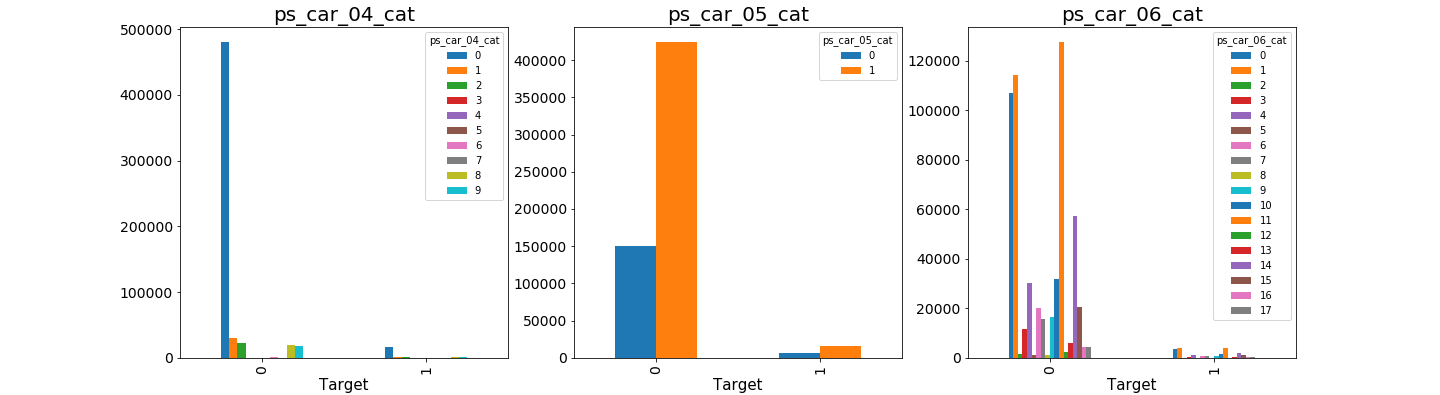
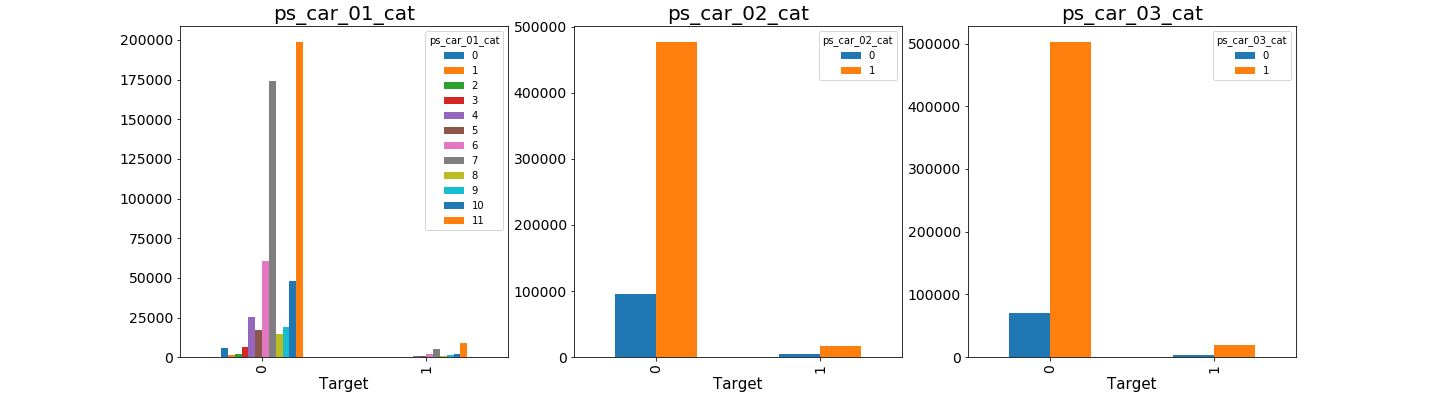
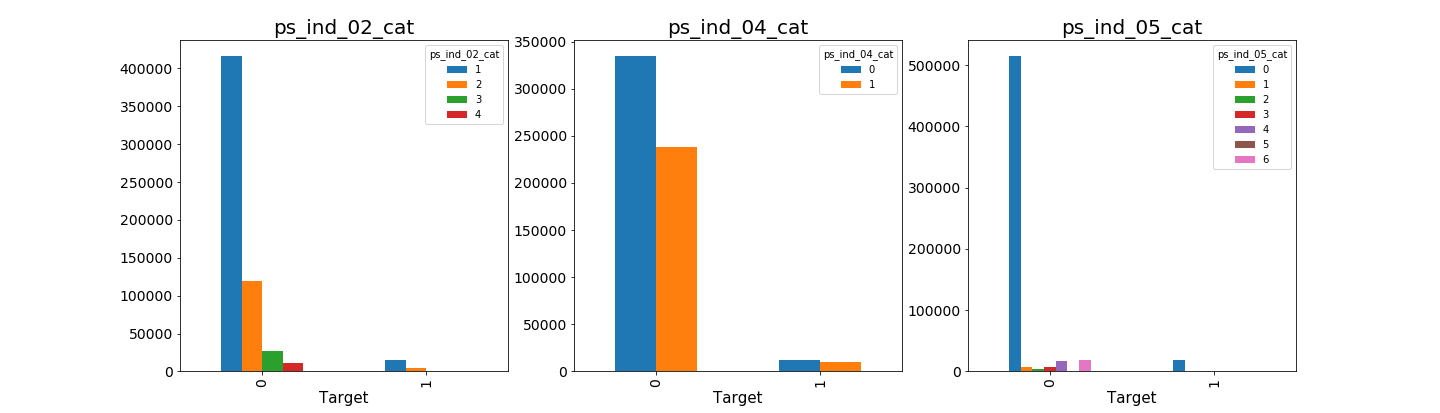
**Figure 1: Class Labels Distribution**



**Table 1: Data Description Table**

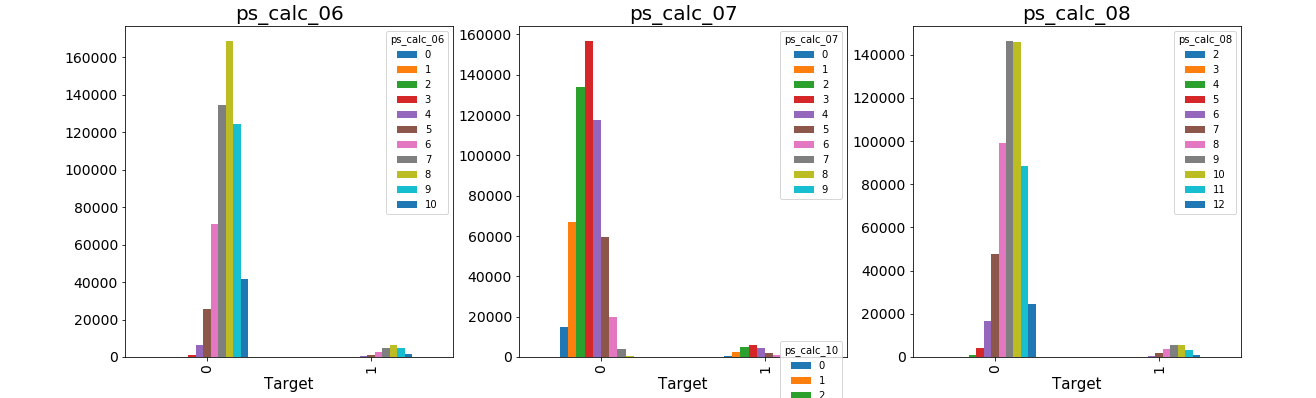
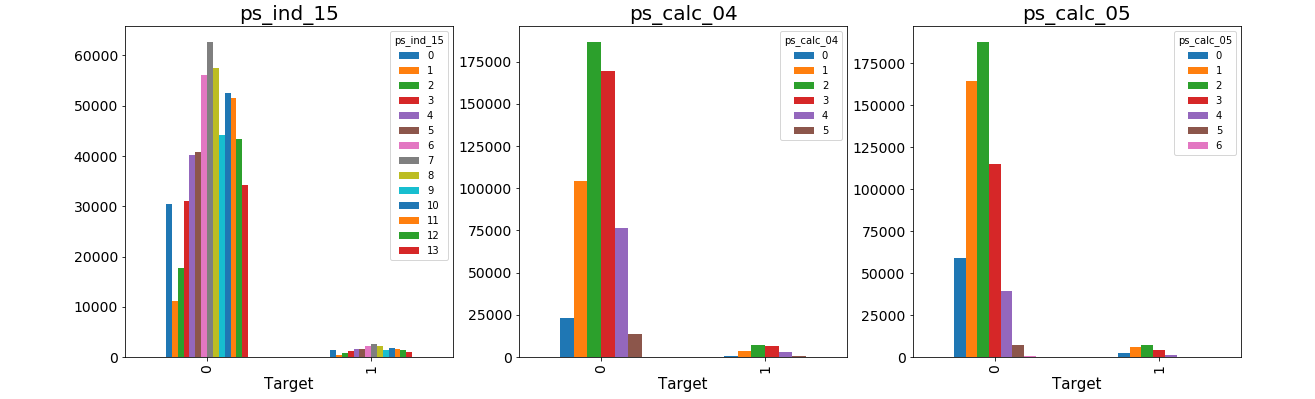
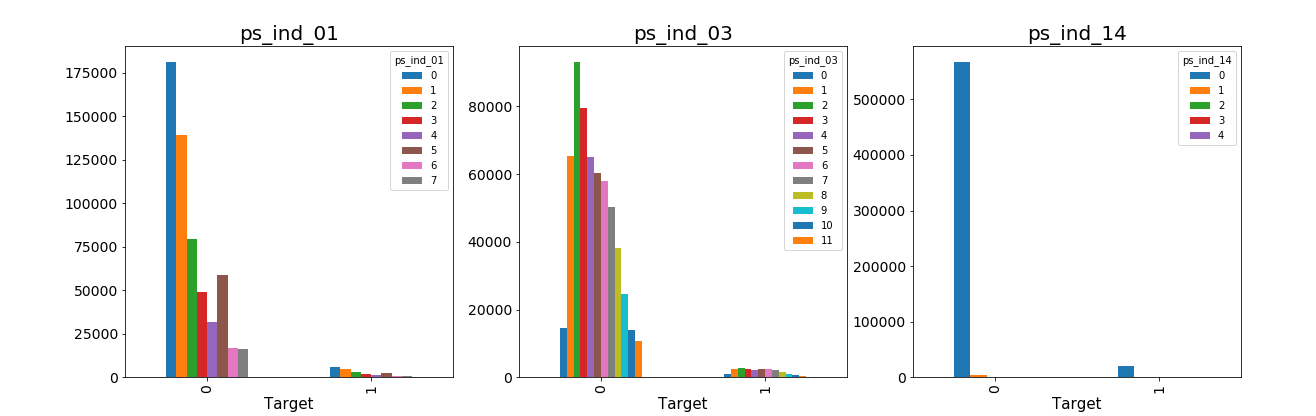
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attributes Name | Data Type | Data Range | Attributes Name | Data Type | Data Range |
| ps\_ind\_01 | Oridinal | 0-7 | ps\_car\_09\_cat | Category | 0-4 |
| ps\_ind\_02\_cat | Category | 0-4 | ps\_car\_10\_cat | Category | 0-2 |
| ps\_ind\_03 | ordinal | 0-11 | ps\_car\_11\_cat | Category | 0-105 |
| ps\_ind\_04\_cat | Category | 0-1 | ps\_car\_11 | Ordinal | 0-3 |
| ps\_ind\_05\_cat | Category | 0-6 | ps\_car\_12 | Continuous | <1 |
| ps\_ind\_06\_bin | Binary | 0 or 1 | ps\_car\_13 | Continuous | <5 |
| ps\_ind\_07\_bin | Binary | 0 or 1 | ps\_car\_14 | Continuous | <1 |
| ps\_ind\_08\_bin | Binary | 0 or 1 | ps\_car\_15 | Continuous | 0-4 |
| ps\_ind\_09\_bin | Binary | 0 or 1 | ps\_calc\_01 | Continuous | 0-1 |
| ps\_ind\_10\_bin | Binary | 0 or 1 | ps\_calc\_02 | Continuous | 0-1 |
| ps\_ind\_11\_bin | Binary | 0 or 1 | ps\_calc\_03 | Continuous | 0-1 |
| ps\_ind\_12\_bin | Binary | 0 or 1 | ps\_calc\_04 | Ordinal | 0-5 |
| ps\_ind\_13\_bin | Binary | 0 or 1 | ps\_calc\_05 | Ordinal | 0-6 |
| ps\_ind\_14 | Ordinal | 0-4 | ps\_calc\_06 | Ordinal | 0-10 |
| ps\_ind\_15 | Ordinal | 0-12 | ps\_calc\_07 | Ordinal | 0-9 |
| ps\_ind\_16\_bin | Binary | 0 or 1 | ps\_calc\_08 | Ordinal | 0-12 |
| ps\_ind\_17\_bin | Binary | 0 or 1 | ps\_calc\_09 | Ordinal | 0-7 |
| ps\_ind\_18\_bin | Binary | 0 or 1 | ps\_calc\_10 | Ordinal | 0-25 |
| ps\_reg\_01 | Continuous | 0.0-0.9 | ps\_calc\_11 | Ordinal | 0-19 |
| ps\_reg\_02 | Continuous | 0.0-1.8 | ps\_calc\_12 | Ordinal | 0-10 |
| ps\_reg\_03 | Continuous | 0.0-3.0 | ps\_calc\_13 | Ordinal | 0-13 |
| ps\_car\_01\_cat | Category | 0-11 | ps\_calc\_14 | Ordinal | 0-23 |
| ps\_car\_02\_cat | Category | 0 or 1 | ps\_calc\_15\_bin | Binary | 0 or 1 |
| ps\_car\_03\_cat | Category | 0 or 1 | ps\_calc\_16\_bin | Binary | 0 or 1 |
| ps\_car\_04\_cat | Category | 0-9 | ps\_calc\_17\_bin | Binary | 0 or 1 |
| ps\_car\_05\_cat | Category | 0 or 1 | ps\_calc\_18\_bin | Binary | 0 or 1 |
| ps\_car\_06\_cat | Category | 0-17 | ps\_calc\_19\_bin | Binary | 0 or 1 |
| ps\_car\_07\_cat | Category | 0 or 1 | ps\_calc\_20\_bin | Binary | 0 or 1 |
| ps\_car\_08\_cat | Category | 0 or 1 |  |  |  |

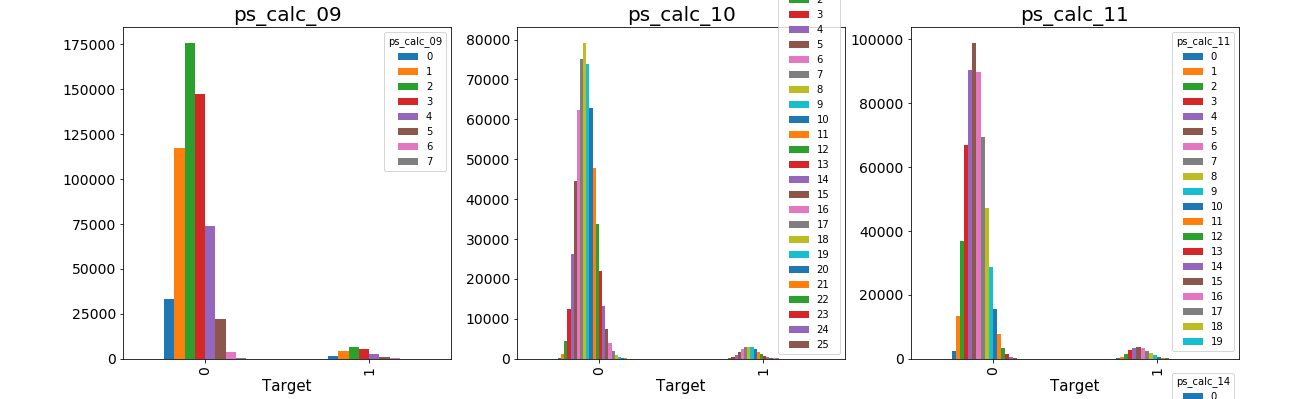
The basic data description on table 1 shows that there are 4 types of variables (i.e. Continuous, Binary, Categorical, Ordinal) in the dataset. The distributions of each type of variables are displaying from figure 2- figure 5 below.

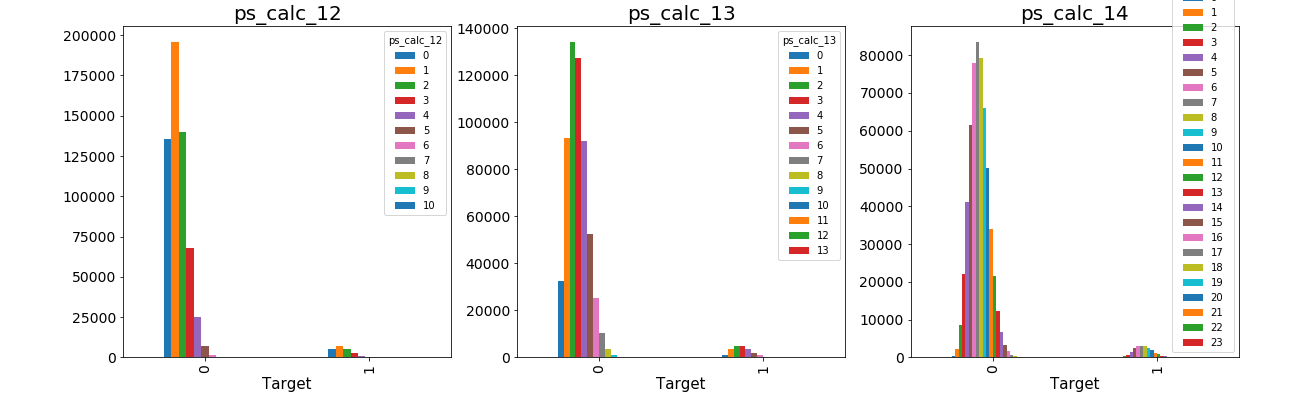
**Figure 2: Categorical Variables plots**

As we mentioned, the dataset is highly biased on the Class “0”. The distribution plot on Categorical variables on figure 2 basically shows all the attributes are highly biased to the class “0” which is not surprised. However, we could also focus more on the attributes ps\_car\_10 cat, ps\_car\_07\_cat and ps\_ind\_05\_cat due to the ratio of these attributes from Class “0” to Class “1” is about 500:1 which means that almost all the class “0” has a vote on these attributes.

**Figure 3: Ordinal Variables plots**

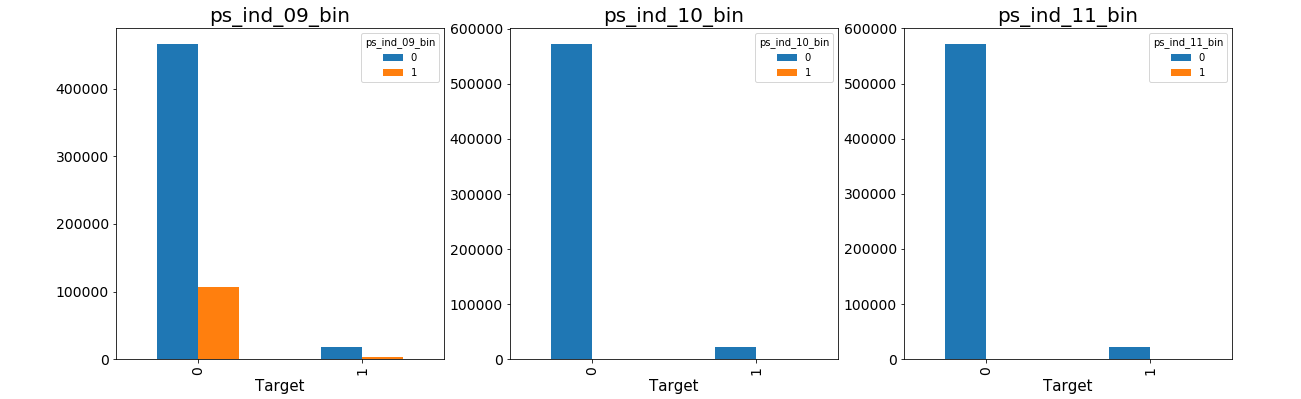
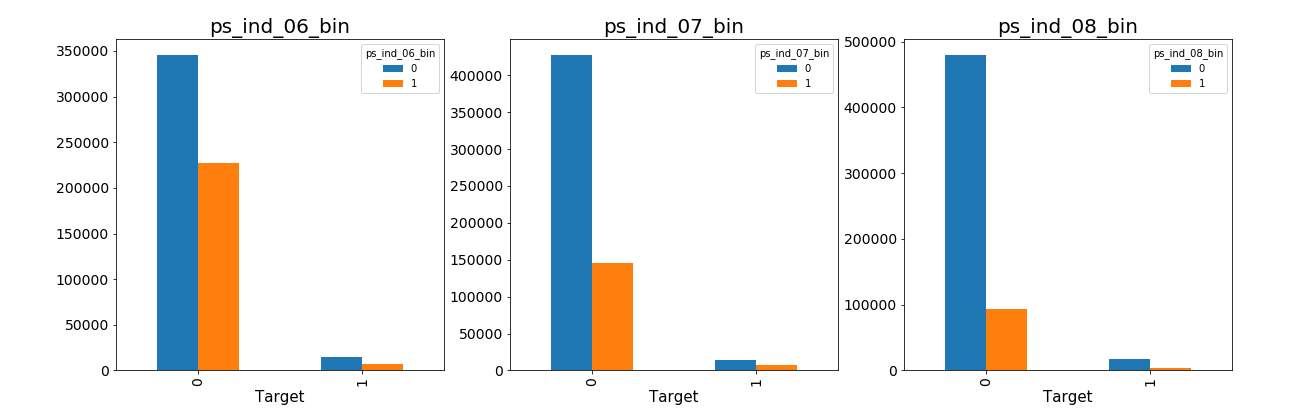


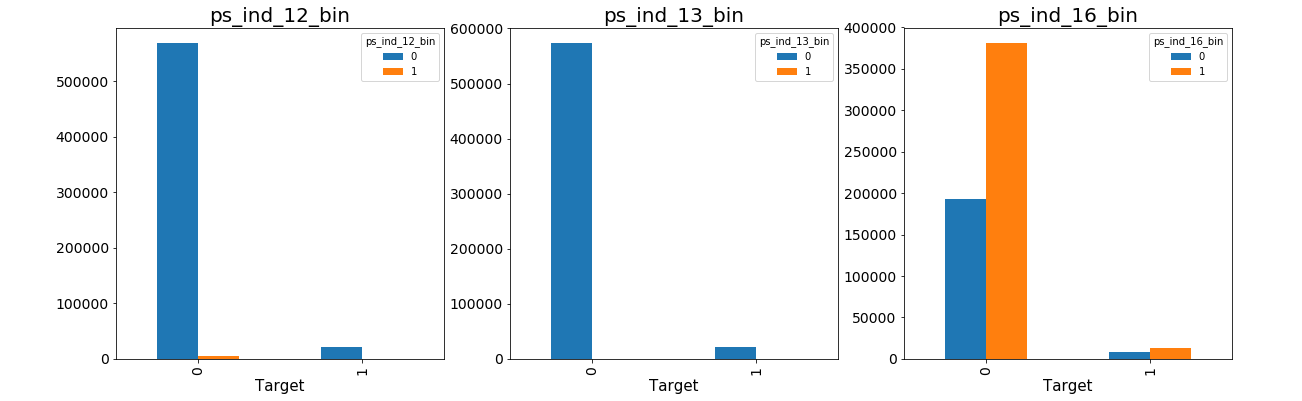


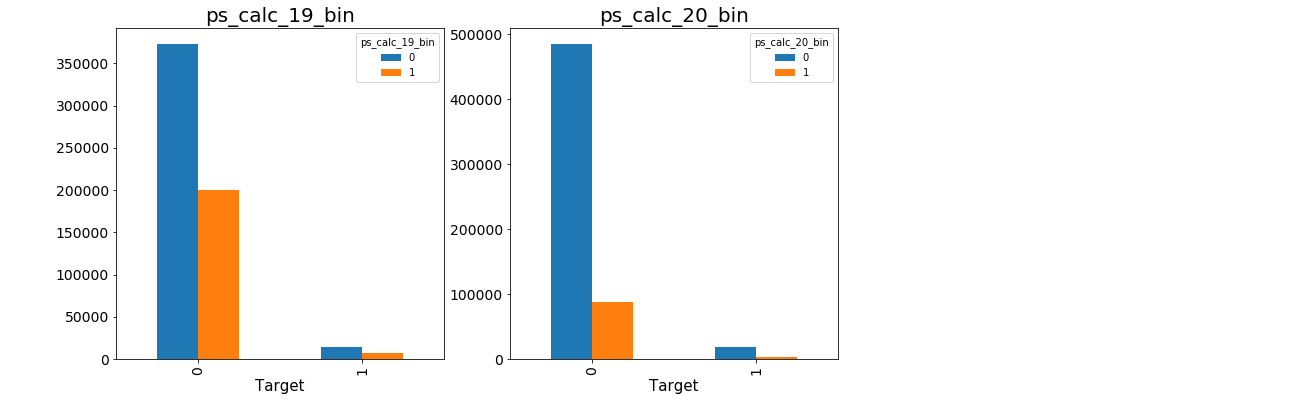
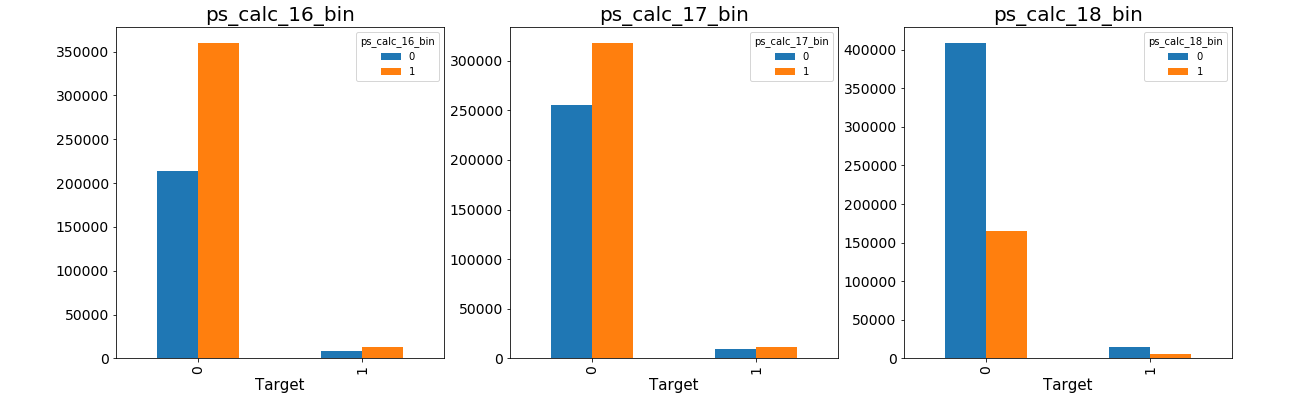
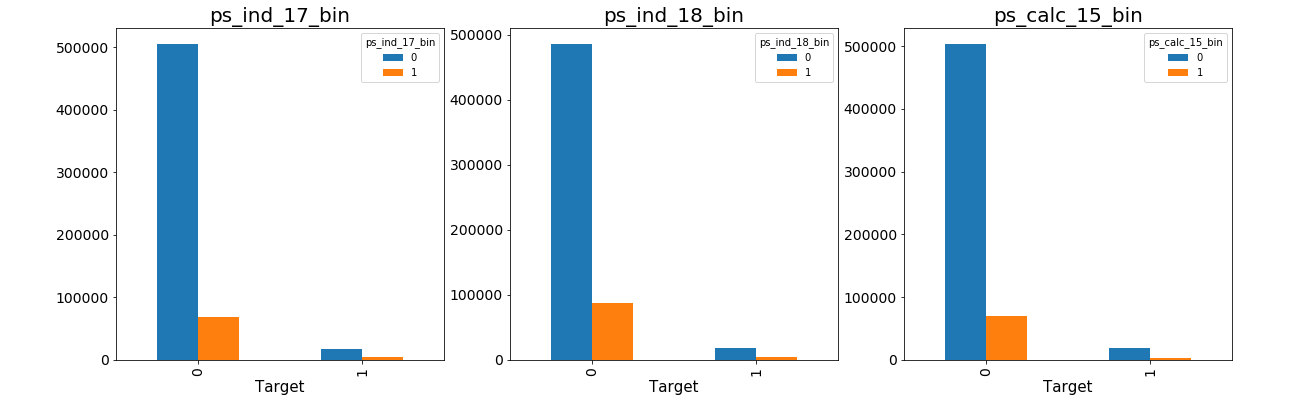


Similar to the distribution of Categorical variables, the plots above on figure 3 also shows all the ordinal attributes are highly biased to the class “0” for this highly biased dataset. However, we could also focus more on the attribute ps\_ind\_14 since the ratio of this attributes from Class “0” to Class “1” is about 500:1 which means that almost all the class “0” has a vote on this attribute. We are able to check whether the final model shows this attribute would be one of the most important features to the model.

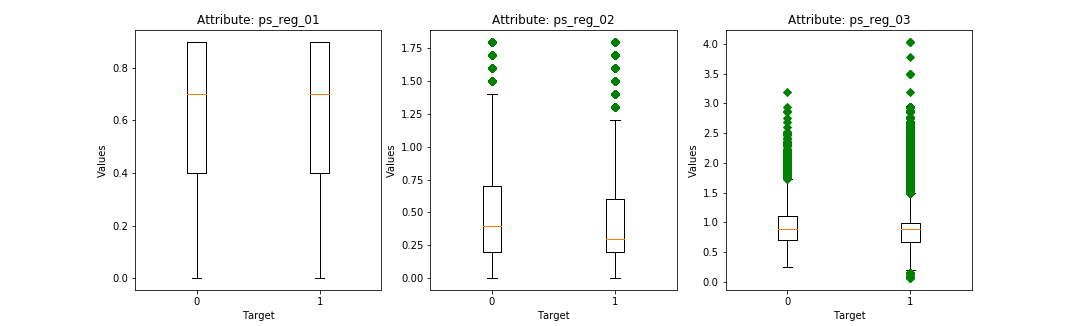
**Figure 4: Binary Variables plots**

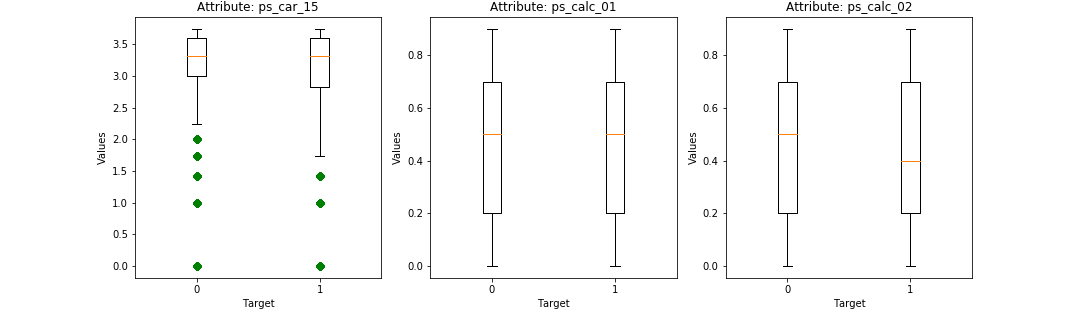
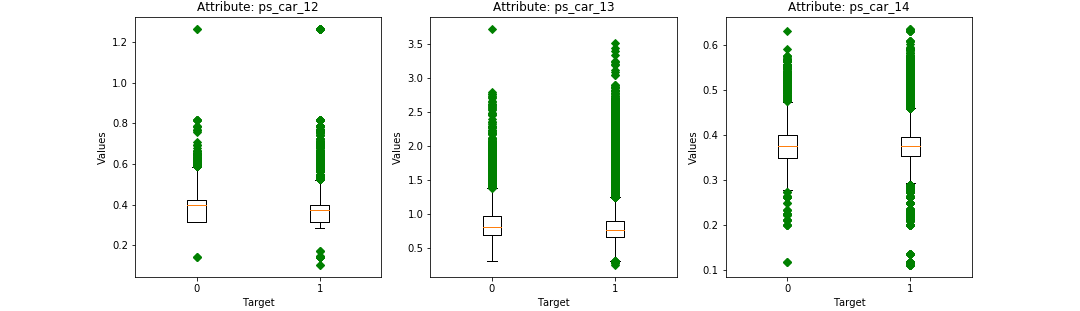






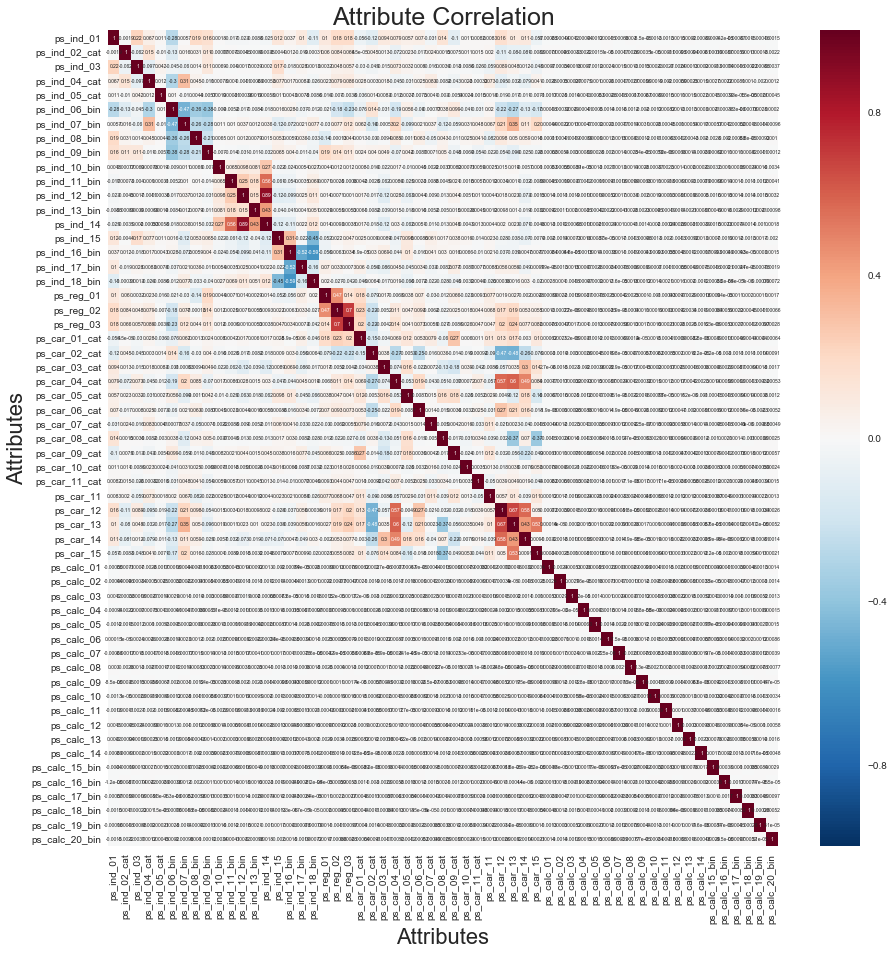
The plots above on figure 4 basically shows all the Binary attributes are highly biased to the class “0” as well. We could also focus more on the attributes ps\_ind\_08\_bin,ps\_ind\_10\_bin, ps\_ind\_11\_bin, ps\_ind\_13\_bin, ps\_ind\_15\_bin, ps\_ind\_17\_bin, ps\_ind\_18\_bin since the ratio of this attributes from Class “0” to Class “1” is more than 500:1 which means that almost all the class “0” has a vote on these attributes. Cross-checking on the final models to find out whether these attributes could be one of important features to the models.

**Figure 5: Continuous Variable plots**



The continuous attributes on figure 5 shows a little different than the categorical / ordinal variables, the range of the continuous variables do not seem to differentiate between the target attribute. However, some of the attributes have more outliners on the side of Class “1”, e.g. ps\_reg\_03, ps\_car\_13, ps\_car12. Because of that, we could observe more if these variables would affect the final model predictions.

**Figure 6: Correlation heat map**



Due to absence of detail variables description, we would not be able to analysis the relation between attributes in very detail. Based on the figure 6, we could see small extent of correlation between attributes. Most of attributes are not correlated with each other except few attributes, e.g. ps\_ind\_14 and ps\_ind\_12\_bin, ps\_reg\_03 and ps\_reg\_02.

**Related Works**

Insurance companies widely use data mining and machine learning techniques to account for risk management and customers’ premium. They attempt all kinds of way to predict the customers’ behavior in order to minimize their risk while estimate the premiums for the customers. The following researches or projects are some examples related to this topic. For example, the reference project #1 by Chu-Shiu Li, Chwen-Chi Liu and Sheng-Chang Peng predicted customers’ auto premiums amount in Taiwan. The project #2 by Dan Huangfu predicted Bodily injury claim amount on auto insurance etc. According to these researches, we found that our experiment has some common challenges in this topic: 1) Dataset are typical in high dimensional. 2) Weak correlation between Target and Predictor variables. 3) Dataset contains large amount of missing values. With these type of challenges, we will attempt several distinctive approaches to tackle the dataset in order to achieve the optimal Accuracy in this classification experiment.

1. Title: Risk Classification and Claim Prediction: An Empirical Analysis from Vehicle Damage Insurance in Taiwan: This project which is titled as”” . The authors are Chu-Shiu Li, Chwen-Chi Liu and Sheng-Chang Peng who are from National Kaohsiung First University of Science and Technology, Feng Chia University and Ming Chuan University respectively.

The aim of this project is to predict premium amount that policyholders should pay based on the features of the policyholders (age, gender, and marital status) and their vehicles (car age, car model, and exhaust) as well as their basic premiums. Also another goal of this paper is to evaluate whether BMS systems are able to monitor or decrease probabilities of filing claims in the future.The dataset is spitted into two sets by their year (2010 and 2011). Each dataset includes three predictions approaches which are New car policies, New policies and Renewed policies are consist of approximately 110000, 160000, 110000 instances respectively.

This project applies the binary classification method which is Logistic regression analysis (Cut-off =0.5). They evaluate the models with holdout samples method and the following measuring metrics: Sensitivity, Specificity and Accuracy. The authors remarked that all the predictor variables are significantly important to the models and response target variable. Model: Premium ＝ Basic Premium × (Gender-age Coefficient + Claim Coefficient) × ManufactureCoefficient

Result:

New Car policy: sensitivity=70-80%, specificity=53-55%, Accuracy=65-69%

New policies: Sensitivity=70-77%, specificity=55-57%, Accuracy = 63-68%

Renewed policies: Sensitivity=40-44%, specificity=83-85%, Accuracy = 68-71%

1. Title: Data Mining for Car Insurance Claims Prediction is published in April 2015. The author is Dan Huangfu who in Worcester Polytechnic Institute. The aim of this project predicts injury claims in 2008from policyholders based on the past 3-year (2005-2007) insurance claims from Allstate Insurance company.

The entire dataset is consist of 13184290 instances and 34 predictor variables. The target variable is the round-up claim amount . Predictor variables include vehicle model year, manufacturer, sub-model etc and the rest of the variables' features are not described. The author attempted to apply multiple methods to run on the prediction including logistic regression, Tweedie’s compound gamma-Poisson model, principal component analysis (PCA), response averaging, and regression and decision trees. The final determined method is PCA combined with a with a Regression Tree which produced the best results .

The author remarked couple of challenges during working on the project as the followings: 1) High dimensional dataset, 2) Weak correlation between Target and Predictors. 3) Dataset contains numerous missing values.

1. Title: Comparision of Data Mining Techniques for Insurance claim Prediction is published in 2010/2011. The author is Andrea Dal Pozzolo. The aim of this project is to predict Bodily injury liability insurance claim payments in dollars.

The information of the dataset is from 2005-2007. There are 32 predictor variables which are the characteristics of the insured customers' vehicle includes continuous, categorical and Integer. Missing values are presented from the original dataset (2mil instances). The author did the project in 2 ways: 1) Predict the continuous values by regression 2) Predict the binary class by classification method (Class =1 means claim values greater than zero, or Class = 0 means No claims will be filed)

Authors start with regression prediction by applying all the original predictor variables to predict the responsive variables. Afterward, author transformed categorical into binary format and performed data dimensional reduction by PCA and k-Mean methods. And then Authors carried out the regression and classification (Decision Tree, R.F., KNN, SVM, LDA etc.) with 10-fold cross-validation by utilizing the reduced data. The final result is considerably good with clustering and PCA due to redundant variables from the original dataset. The best result came out from Single Decision Tree Model with PCA which obtains the highest Gini Index values.

The author remarked couple of challenges during working on the project as the followings: 1) High dimensional dataset, 2) Handling categorical variables for regression.

1. Title: Data Mining to Predict and Prevent Errors in Health Insurance claims Processing. The authors are Mohit Kumar, Rayid Ghani, Zhu-Song Mei who are from Accenture Technology Labs. The aim of this project is to predict the filed medical insurance claims would need to be re-worked due to process errors.

The sourced data was captured through the Claim Processing Pipline. The features of the dataset are based on Member information, Provider information, Claim Header and Claim Lline Details. The number of features was nearly 110,000 binary features (Divide from category features). The Target variable is in binary format which results in Class =1 meaning payment needs to be reworked, or otherwise.

Authors used one of the SVM package which is called SVMperf to handle the large dataset because SVMperf runs fast in the training stage. Besides, Frequency-based feature selection techniques was used to reduce the data dimenision down to 8200 features. During the analysis, authors used Precision (or hit rate) and Recall to measure the top 5-10% of claims/documents. The estimated recall rate from this experiment is about 32% (42% from experiment - 10% from industry ) more accurate to predict if the claims are needed to rework. The stated difficulty of handling this experiment is how to efficiently running the model with approximately 200k claims every day.

1. Title: Comparative analysis of machine learning techniques for detecting insurance claims fraud was published in 2015. The authors are R Guha - Head of Corporate Business Development., Shreya Manjunath – Product Manager and Kartheek Palepu – Associate Data Scientist from Apollo Platforms Group. The goal of this project is to predict the insurance fraudulent cases which filed for the favor of individual based on incorrect / irrelevant information. According to the paper, FBI estimates the cost of fraudulent cases is about 40 billion a year.

The dataset was divided into 4 sub-sets which possess different characteristics (i.e. Multiple parties involve claim, Unknown drivablility status on vehicle, claims was not reported to police, Old vehicles claims). The total instances of this project is about 1.2 million.

Authors start off cleaning the data (i.e. eliminate missing values and duplicates). Prior to building the models, features selection analysis was carried out included Forward selection, backward elimination and PCA. Finally, authors constructed multiple M.L. Models including Logistic Regression, Modified Multi-variate Gaussian, Boosting and Bagging using Adjusted Random Forest to compare the results.

The out-performed model was Logistic Regression in terms of Recall, Precision measurement metrics in all 4 datasets. The best recall score was 92% and the best Precision was 100% among all 4 datasets. The challenges in this project includes class imbalance, missing values, handling categorical attributes, bad data quality and Data Errors (e.g. instances duplicated)

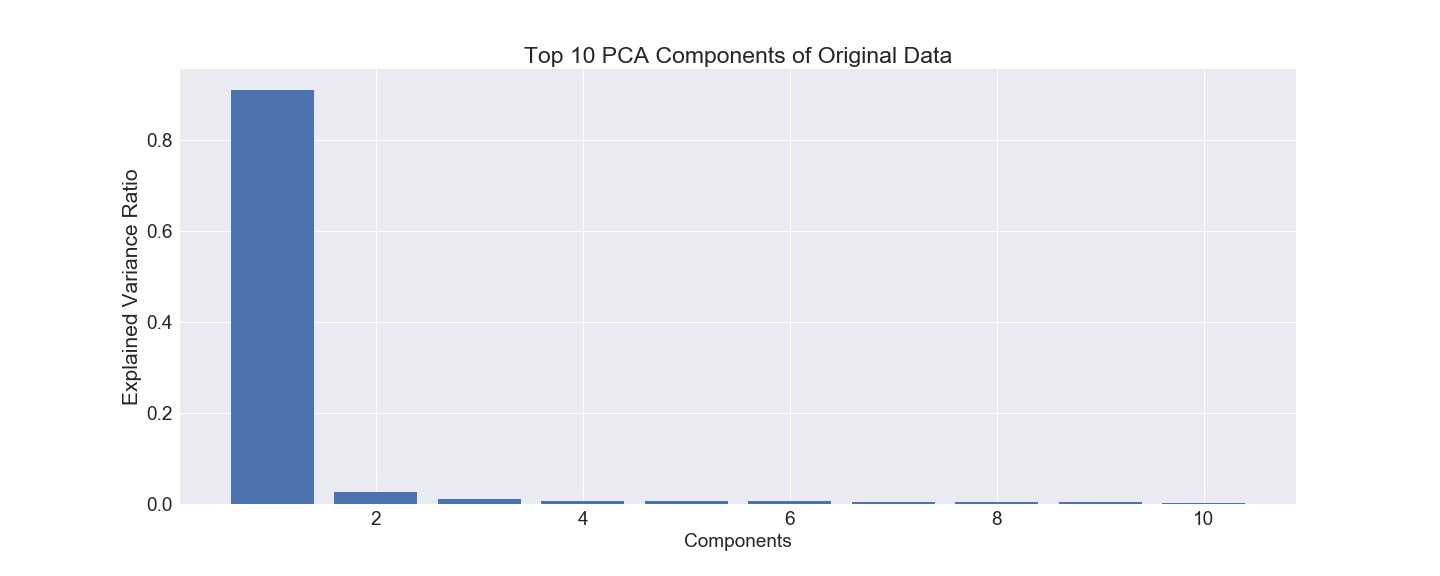
Title: Improving Naive Bayes Models of Insurance Risk by Unsupervised Classification” is published in the textbook ” 2008 International Multiconference on Computer Science and Information Technology”. The authors are Anna Jurek and Danuta Zakrewska from Technical University of Lodz. The goal of this project is to use data mining techniques to evaluate the risk connected with customers who has Life insurance.

The dataset was artificially generated based on research which is not provided by related insurance companies. The target classe was splitted into 3 classes: 1) Customers contains Low risk (Best customers) 2) Customers contains Medium risk 3) Customers contains highest risk (Worst customers). Their best desired aim for this project is to identified the “Best Customer” group. There are total eleven predictor attributes which are transformed form numerical values into Categorical variables.

In this experiment, there are two approaches were implemented: 1) Classification by Naives Bayes model based on the entire dataset. 2) Clustering was carried out based on two sub-set of data which was obtained from prior feature selection process. The final result showed that the clustering approaches with sub-set of data out-performed than only classification on entire set of data. The highlighted methodology is that using feature selection prior the the learning process would increase the efficiency and accuracy of the classification result.

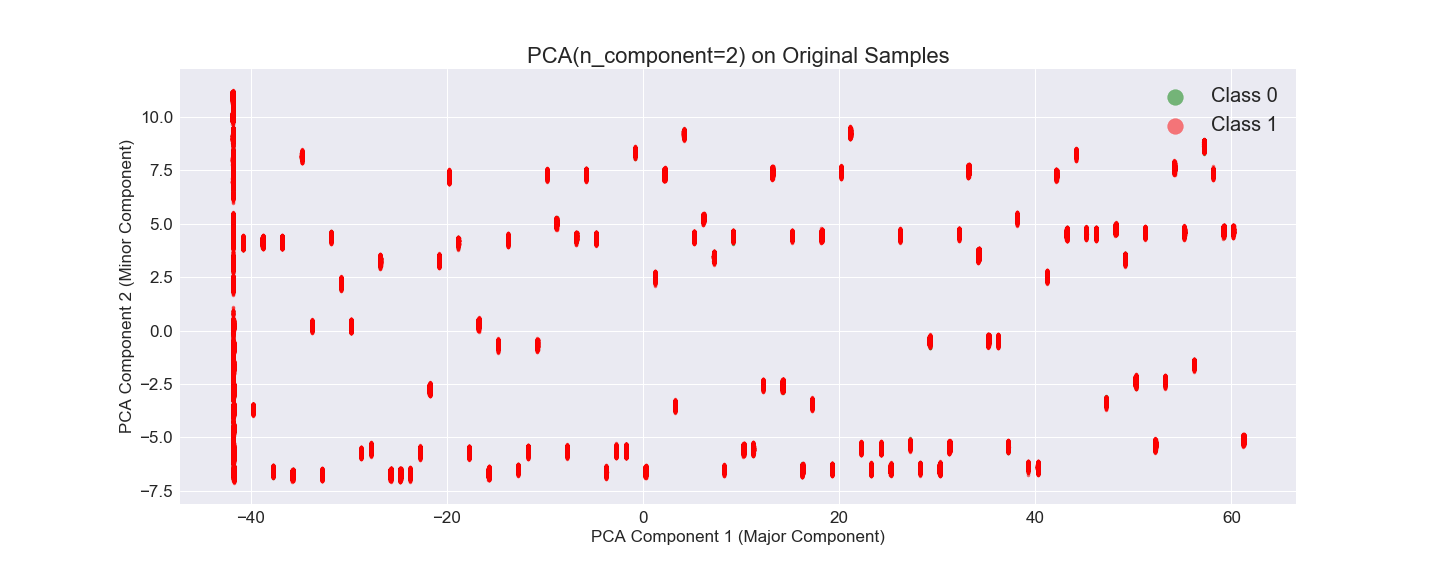
**Methodology:**

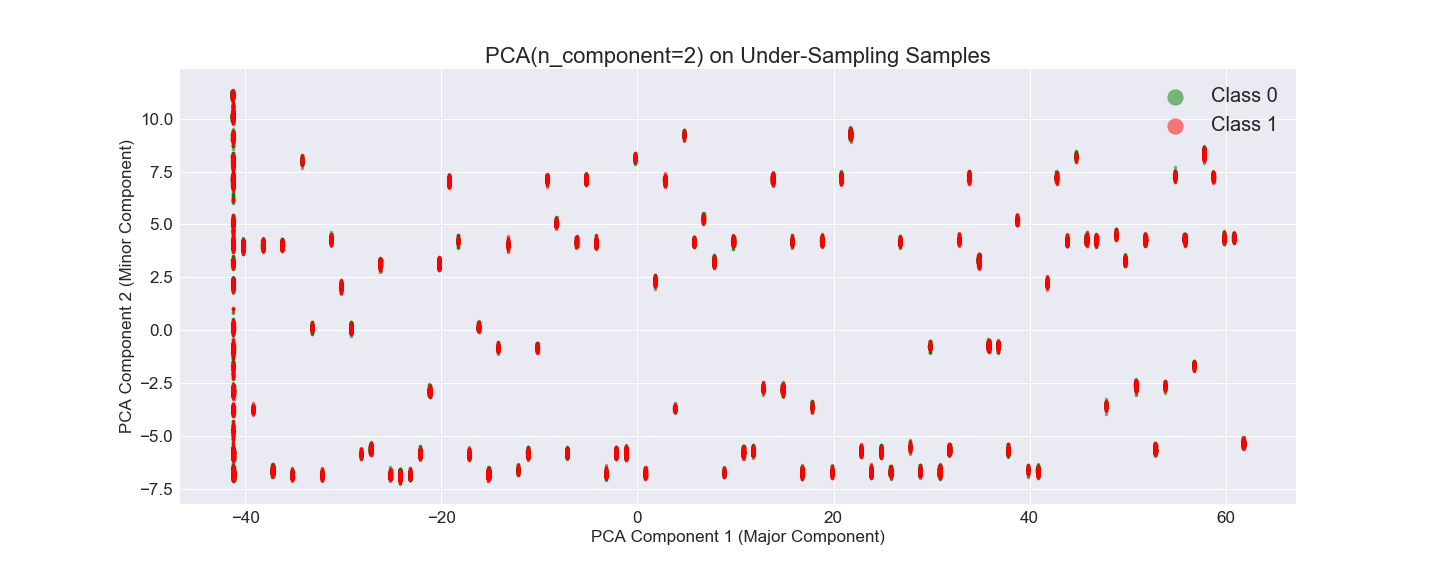
The primary tool was exploited in this project is Python software with Scikit-Learn library. This experiment was first started with the data pre-processing steps, we have replaced the missing values (values=-1 in this case) with the mean values for continuous variables and the mode values for the categorical and ordinal variables. With this large-scale dataset, the next step we performed the feature extraction by PCA. By observing the figure 7 below, we have realized that the first component represented more than 90% variation of the target variable among all the variables.

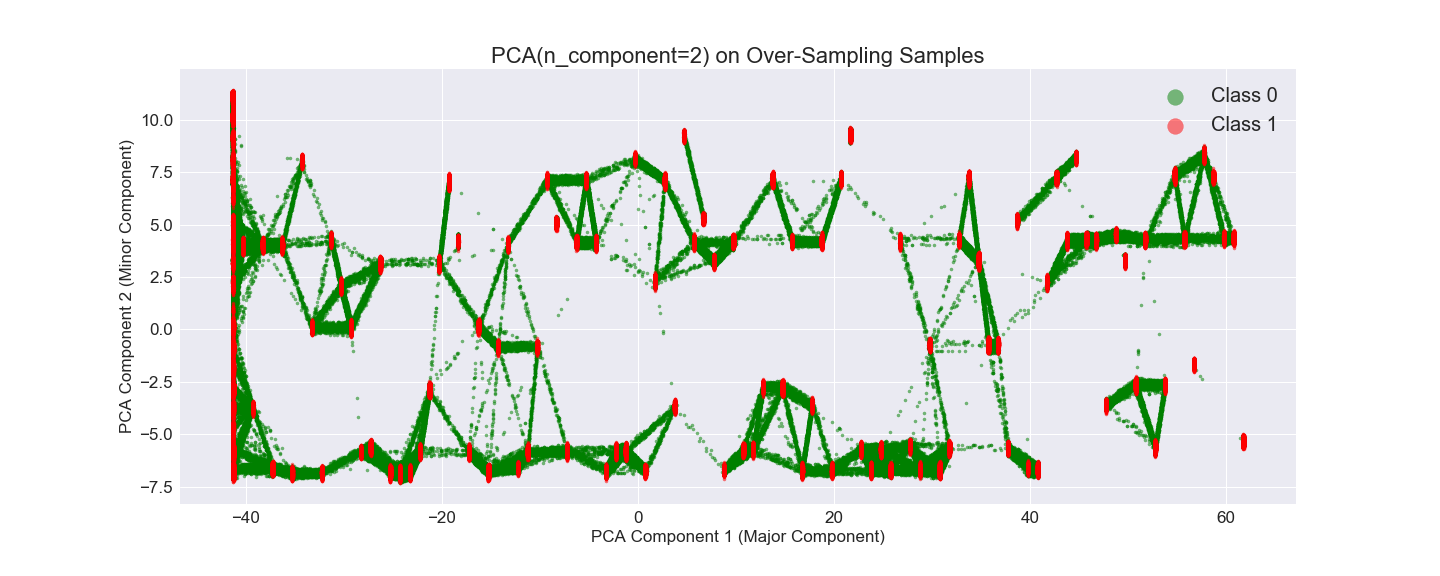
Figure 7: Top 10 PCA Components on Original Data

The original dataset is consisted of large imbalanced data which is bias in favor the majority class = “0”. The ratio of the class “0” to the ratio of class “1” is approximate 26:1. In the original dataset PCA scatter plot shown on figure 8, the binary classes in the original dataset are not partitioned in a clear separated space. In order to experiment whether the outcome is related to the Class distribution, over-sampling and under-sampling techniques were applied to the data mining process. The followings are the PCA(Components=2) plot visualizing the distribution of the binary Classes.

**Figure 8: PCA Plot on Original Samples**



**Figure 9: PCA Plot on Under-Sampling Data**

**Figure 10: PCA Plot on Over-Sampling data**

As we can see above, the under-sampling plot (figure 8) is very similar to the pattern of the original data; the over-sampling plot (figure 9) is more distinguishable than both original and under-sampling data. By viewing that, we could assume that over-sampling data might offer a better classification result.

Several of data mining and machine learning algorithms are employed to this experiment. We started on testing on two single classifiers (i.e. Decision Trees, Naïve Bayes) and three ensemble classifiers (i.e. Random Forest, Bagging, AdaBoost) on the above three format of data samples. In order to achieve our goal which is to accurately find out whether the policyholders would file the claims in the next year, various measurement metrics which are Accuracy, Recall, Specificity, Precision are utilized. Cross validation and holdout sampling techniques are applied and results are listed as shown below.

**Experimental Result:**

**The sequence of this experiment is to test on the Original, Under-Sampling (21694 instances on each class) and Over-Sampling (**573518 instances on each class**) dataset respectively by exploiting the mentioned FIVE classifiers. As the domain of this project goes, our focus is to optimize the Sensitivity / Recall score (i.e. Predict policyholders would file claims in the following year). The experiment is carried out in a 3-fold Cross-Validation (C.V.) format and the results are shown as below. The chosen ensemble classifiers’ results are based on the optimal scores with the least complexity which means that the more number of estimators, the more complex of the model. The scores of Specificity, Sensitivity and Precision are calculated based on the Validation set of data.**

**Experiment Result: Original Samples (Table 2):**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **CLASSIFIER** | **TRAIN SCORE** | **TEST SCORE** | **SPECIFICITY** | **SENSITIVITY** | **PRECISION** | **C.V.** |
| DECISION TREE | 100% | 92.3% | 95.6% | 5.79% | 4.71% | 3 |
| BERNOULLINB | 96.4% | 96.4% | 100% | 0.0% | 0.0% | 3 |
| BAGGING -100 Est  (DECIOSN TREE) | 100% | 96.4% | 100% | 0.0% | 0.9% | 3 |
| ADABOOST -100 Est  (DECISION TREE) | 96.4% | 96.4% | 100% | 0.0% | 0.0% | 3 |
| RANDOM FOREST  100 Estimators | 100% | 96.4% | 100% | 0.0% | 0.0% | 3 |

**According to the above table 2, the accuracy of the original sample which are all above 90% are in an acceptable level. Also, the accuracy of training and testing set are very closed which implies over-fitting issue would not be the case. However, the Sensitivity scores are very low among all the classifiers, most of scores equals to ZERO except Decision Tree model. As we discussed above, this low recall score might be leaded by the imbalanced original data. Now, let’s attempt on the Under-Sampling experiment.**

**Experiment Result: Under-Sampling (Table 3):**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **CLASSIFIER** | **TRAIN SCORE** | **TEST SCORE** | **SPECIFICITY** | **SENSITIVITY** | **PRECISION** | **C.V.** |
| DECISION TREE | 100% | 52.3% | 52.4% | 52.2% | 52.3% | 3 |
| BERNOULLINB | 57.6% | 57.5% | 61.1% | 53.8% | 58.1% | 3 |
| BAGGING -100 Est  (DECIOSN TREE) | 100% | 58.1% | 60.9% | 55.3% | 0.9% | 3 |
| ADABOOST -100 Est  (DECISION TREE) | 100% | 60.3% | 59.2% | 61.2% | 57.2% | 3 |
| RANDOM FOREST  100 Estimators | 100% | 58.2% | 60.5% | 55.9% | 58.6% | 3 |

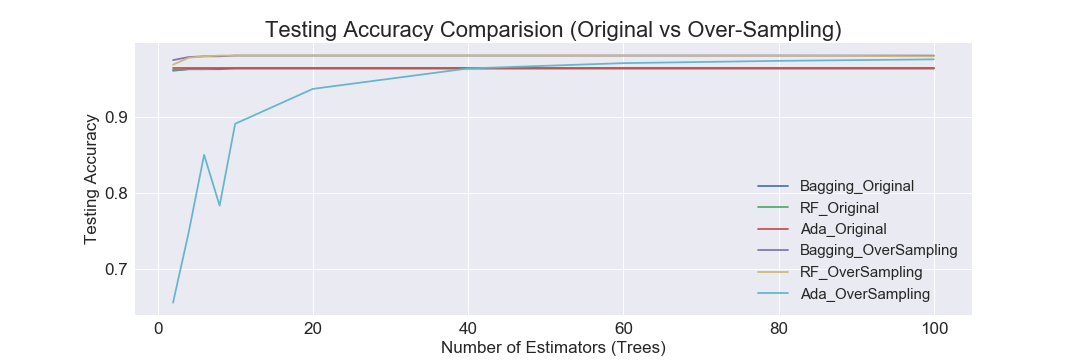
**The Under-Sampling dataset leads to boost the Sensitivity score higher among all the classifiers as shown on Table 3. However, the accuracy on the testing set dramatically drops down around 55%-60% level. In this case, over-fitting issue between the training and testing dataset is probably the cause. Also, the specificity drops along with accuracy as well. So that this result obviously not able to achieve the goal of this project. Let’s attempt on the Over-Sampling experiment.**

**Experiment Result: Over-Sampling (Table 4):**

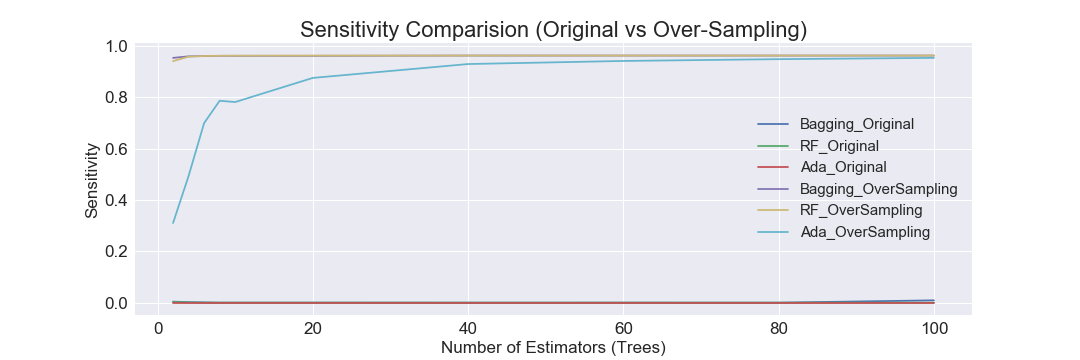
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **CLASSIFIER** | **TRAIN SCORE** | **TEST SCORE** | **SPECIFICITY** | **SENSITIVITY** | **PRECISION** | **C.V.** |
| DECISION TREE | 100% | 95.8% | 95.4% | 96.25% | 95.5% | 3 |
| BERNOULLINB | 85.3% | 85.3% | 83.9% | 86.7% | 84.3% | 3 |
| BAGGING - 6 Est  (DECIOSN TREE) | 100% | 98.1% | 100% | 96.1% | 100% | 3 |
| ADABOOST - 6 Est  (DECISION TREE) | 97.6% | 97.6% | 100% | 69.9% | 100% | 3 |
| RANDOM FOREST  6 estimators | 100% | 98% | 100% | 96.2% | 100% | 3 |

**The above Over-Sampling techniques offers large amount of rise on most of the classifiers except the AdaBoost Classifiers. Since the above result on the ensemble classifiers are based on the optimal result with the least complexity, so that we only chose the ensemble classifiers with SIX estimators (i.e. Number of trees). The Over-Sampling technique performs good on single Decision Tree, Bagging with Decision Tree and Random Forest Tree Classifiers which offers above 95% scores on Accuracy and Sensitivity.**

**Figure 11: Testing Accuracy Comparison**



**Figure 12: Sensitivity Score Comparison**



We wanted to see how the number of trees effected on the ensemble models. As seen above figure 11, the accuracy of all ensemble models are basically above 95% at around 40 trees. Also, figure 12 shows that both bagging and random forest require only few trees to reach maximum sensitivity. Adaboost maxes out at around 40 trees. When it comes to using the original samples, it is impossible to improve the sensitivity regardless of the amount of tree we use. Because of the computational expense of this dataset, we would like to focus on bagging and random forest as they will take the least amount of time to train.

**Model Comparison (Table 8):**

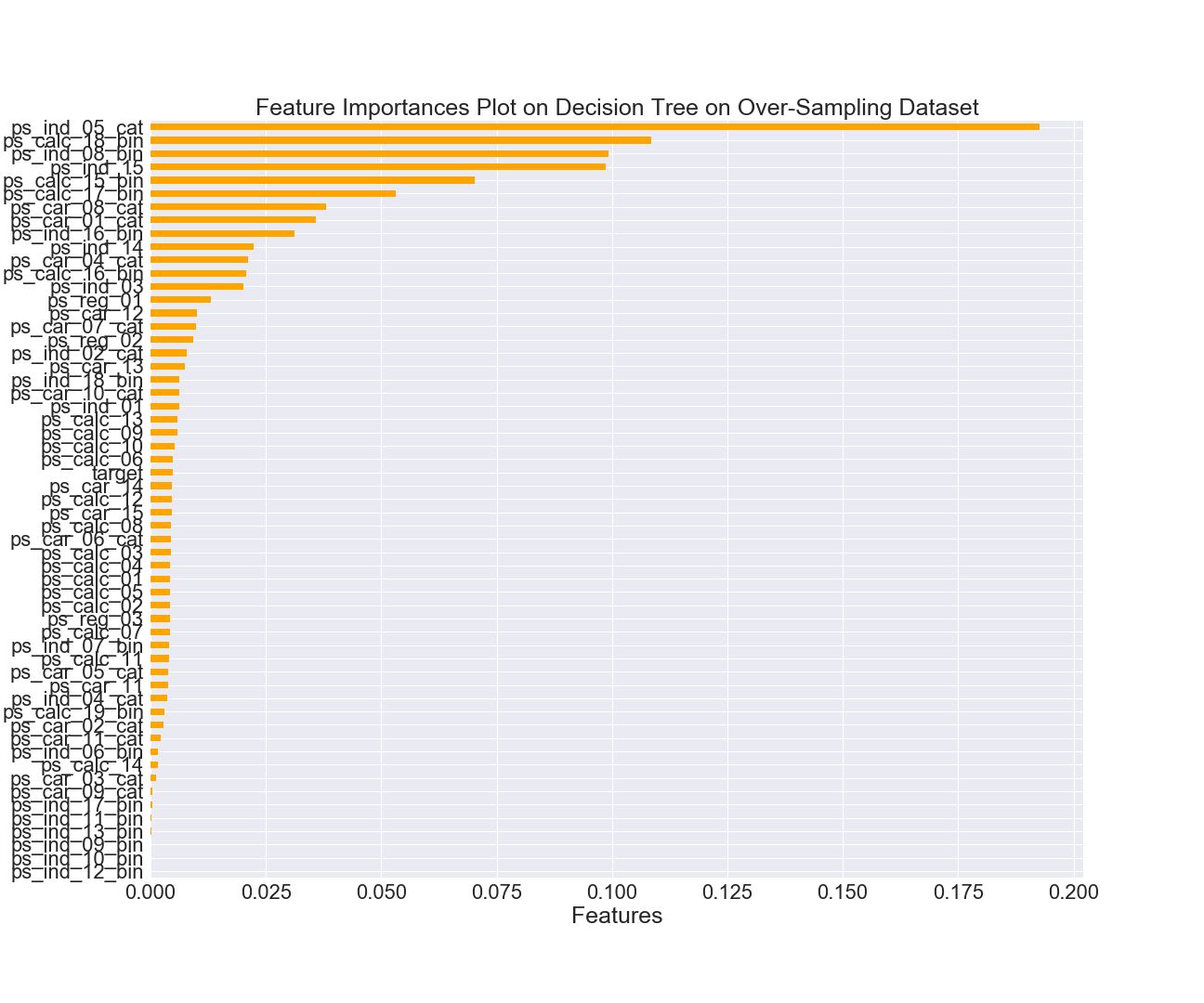
|  |  |  |  |
| --- | --- | --- | --- |
| **MEAN VALUE**  **(20 holdout Test)** | **DECISION TREE** | **BAGGING(Est =6)** | **RANDOM FOREST (Est=6)** |
| **SENSITIVITY** | 96.2% | 96% | 96% |
| **Confidence Interval (Sensitivity)** | (87.9%, 104.6%) | (87.5%,104.6%) | (87.5%,104.6%) |
| **ACCURACY** | 95.8% | 98% | 98% |
| **Confidence Interval (Accuracy)** | (87%, 104.6%) | (91.8, 104.1%) | (91.9%, 104.1%) |
| **Mean Runtime per Holdout test** | 41sec | 174 sec | 165sec |

**Paired t-test (Table 9):**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **P-Value**  **(20 Holdout Test)** | **Decision Tree** | **Bagging** **(Est =6)** | **Random Forest (Est =6)** |  |
| **Decision Tree** |  | 0.000 | 0.000 | **ACCURACY** |
| **Bagging** **(Decision Tree)**  **(Estimators =6)** | 0.000 |  | 0.082 |
| **Random Forest**  **(Estimators =6)** | 0.000 | 0.883 |  |  |
|  | **Sensitivity** |  |  |

To statistically compare our models, we completed a paired t-test by implementing 20 rounds of holdout test on bagging, random forest, and decision trees. In terms of Accuracy and Sensitivity, Single Decision Tree classifier is statistically significant different from ensemble classifiers Bagging and Random Forest as shown on Table 9. In other words, single Decision Tree model statistically out-performed Bagging and Random Forest in Sensitivity metric shown on figure 8. The margin of errors for 95% confidence interval of sensitivity of Decision Tree is not too high which is above 87%. So that I would propose Decision Tree is a good classifier for predicting if drivers would file a claim in the following year (Class = 1). To account for Accuracy result for complete dataset, Bagging and random forest models would be the right choice statistically. The error margin of accuracy in 95% confidence interval of both ensemble classifiers which are lower than single Decision Tree are above 90%. In my opinion, we should evaluate the all three classifiers when new data comes in needed to be evaluate because their accuracy are very closed.

**Figure 13: Feature Importances**



**The Importance of Features on figure 13 shows that the top five features on the Over-Sampling Decision Tree model are categorical, ordinal, binary attributes. The top 4 out of 5 features were observed earlier that having a large ratio gap between Class 1 and Class 0 which proves that Data Analysis steps are an indispensable step for classification process. We could also observe that continuous variables do not offer very great effect on the model as we proposed earlier due to outliers and no obvious dissimilarity on attributes values amid target classes.**

**Discussion**

**The final classification result above is very closed to achieve the goal of this experiment. The primary goal is to predict the policyholders are likely to file a claim in the next year based on a set of features. Secondly, figuring out a reliable predictive model offers high Accuracy and Sensitivity classification outcome.**

**Based on our experiment result, we are able to understand more on how the capabilities of classifiers as well as the sampling techniques affect prediction. There are some common features among the classifiers, single and ensemble, except the Naïve Bayes. The performance of Naïve Bayes classifier is inferior to other classifiers in the way of applying with the Under-Sampling and Over-Sampling dataset. The accuracy and sensitivity of other four models scores increase as the over-sampling population increase, the sensitivity achieves above 96% and the accuracy also achieve above 95%. The maximum of parameters we have attempted on ensemble classifiers (Bagging and Random Forest) is 100 number of estimators.**

**In terms of model evaluation, users always consider various aspects (i.e. Model stability, Model complexity, Prediction Accuracy etc.). In this experiment, we experienced some of the important steps which are required to take into account for model evaluations, e.g. Model complexity and Sensitivity. The domain of this project is to find the positive class target (i.e. Driver will file a claim next year), we certainly need to maximize the “True Positive” score. During the process, we encountered another issue which is computation power which is directly related to the model complexity. The more complex model takes more time to obtain outcome. The best 2 ensemble classifiers, Bagging and Random Forest, both take about 4-fold of time than the single Decision Tree for obtaining the outputs. As we know, the above two ensemble classifiers are constructed with more trees to improves the variance of the outcome. Based on our analysis approach, I would recommend to start on implementing Decision Tree with Over-Sampling method for obtaining the preliminary accuracy on the classification prediction because it takes less time than both ensemble classifiers. After that, implementing the ensemble classifiers for further analysis in order to achieve the higher accuracy and recall scores outcome.**

**Conclusion and Future work:**

We knew going into this analysis that we needed to focus on the sampling techniques that allowed the model to maximize sensitivity. Over-Sampling and Under-sampling were our main techniques at achieving this. Of those, Over-Sampling gave us the best results for sensitivity. With more time and computational power, we would like to take the over-sampled dataset and then focus on tuning other hyper-parameters to further improve the model. Refer to the research by R Guha (Reference 5 above), they have divided the dataset into subsets based on their characteristic and types prior to implement them into the classifiers and each subset might able to provide some forms of result which boosts up the overall accuracy. So that, if we were able to find more description on the original dataset, we would be able to focus more one feature reduction to see how it effects the model performance.

**There are numbers of challenges when we initial the pipeline to analyze this dataset in order to maximize the accuracy, precision etc. Most of the predictor variables are in categorical and ordinal which are not applicable to the binary logistic regression. Also, the dataset is in a fairly large scale with a portion of missing values variables. Based on these challenges, we were able to research out the best model by comparing multiple classifiers by cross-validation and holdout testing. Nowadays , Insurance companies promote plenty of insurance products in the market, the more precision research we could bring in, the more advantages the customers and insurance companies could gain in the long run prospective.**

**References:**

**[1] Guha, R; Manjunath, Shreya; Palepu, Kartheek. (2015).** <http://www.wipro.com/documents/comparative-analysis-of-machine-learning-techniques-for-detecting-insurance-claims-fraud.pdf>

[2] Kirlidog, Melih, Asuk, Cuneyt (2012).

<https://ac.els-cdn.com/S1877042812036099/1-s2.0-S1877042812036099-main.pdf?_tid=37142d16-18f6-11e8-8c65-00000aacb35f&acdnat=1519430820_31a1055eb17a828c75f82c4fadbddd67>

**[3] Mirijanian, Nick; Litven, Joshua. (2016).**

<https://nycdatascience.com/blog/student-works/optimizing-machine-learning-algorithms-model-allstate-loss-claims/>

[4].Chu-Shiu Li, Chwen-Chi Liu, Sheng-Chang Peng, 2013, Risk Classification and Claim Prediction: An Empirical Analysis from Vehicle Damage Insurance in Taiwan, <http://www.wriec.net/wp-content/uploads/2015/07/4G4_Li.pdf>

[5] Dan Huangfu, 2015, Data Mining for Car Insurance Claims Prediction, <https://pdfs.semanticscholar.org/3e9b/cf8d7917b7674d1fdf046c5547c6dc87975b.pdf>

[6]Andrea Dal Pozzolo , 2011, Comparision of Data Mining Techniques for Insurance claim Prediction, <https://pdfs.semanticscholar.org/6273/a3eb59950a28af7bfa81a9d792781cdb2b27.pdf>

[7] Mohit Kumar, Rayid Ghani, Zhu-Song Mei, 2010, Data Mining to Predict and Prevent Errors in Health Insurance claims Processing, KDD, <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.169.1518&rep=rep1&type=pdf>

[8] R Guha, 2015, Comparative analysis of machine learning techniques for detecting insurance claims fraud, <http://www.wipro.com/documents/comparative-analysis-of-machine-learning-techniques-for-detecting-insurance-claims-fraud.pdf>

[9] Anna Jurek and Danuta Zakrewska, 2008, Improving Naive Bayes Models of Insurance Risk by Unsupervised Classification, 2008 International Multiconference on Computer Science and Information Technology, <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.514.2792&rep=rep1&type=pdf>

**Peer Review for final project member: Tyler Jewell**

**Result: 8 out of 10**

**Final Presentation: 8 out of 10**

**Final Report: 7 out of 10**