MA616 Project: Visualize and Predict Potential Customers of Caravan Insurance Policies

Soumil Kanwal
Indian Institute of Technology, Ropar
Punjab, India
2017csb1113@iitrpr.ac.in

1 INTRODUCTION

Reaching out to everyone regarding a company's insurance policy is a very ineffective way of marketing. By identifying the group of customers who are more likely to buy the company's insurance policy, we can very easily reduce a lot of marketing expenses and might come up with much more favorable insurance policy by understanding the key demographic. CoIL ran a challenge in 2000 to better identify and visualize the potential customers for a caravan insurance policy. The dataset contains data of more than 5000 customers. The dataset is based on 85 variables including socio-demographic patterns and product usage data, that can be used to identify whether a person would buy the caravan insurance policy or not.

There are two tasks

(1) Predict potential customers
(2) Describe the potential customers

2 DATASET EXPLORATION

2.1 Data Imbalance

The dataset contains records of 5822 people of which only 348 people buy the insurance policy. It can be inferred that the dataset is highly unbalanced, having less than 10% of data as true positives. We will use Synthetic Minority Oversampling Technique (SMOTE)[1] to handle the class imbalance problem. As shown in [1] we can get better results by oversampling using synthetic data and undersampling the majority class. We will compare the results on Caravan dataset after predictions on original dataset and oversampling with SMOTE.

2.2 Correlation

We can observe from Fig.9 that some variables are highly correlated meaning a rise in one column will trigger rise in another. We will identify them by setting a threshold of 0.5 on absolute value of correlation coefficient.

We can observe that:

(1) average household size *"MGEMOMV"* and household with children *"MFWEKIND"* have a correlation of 0.79
(2) customer main type *"MOSHOOFD"* and customer sub-type *"MOSTYPE"* have a correlation of 0.99. This relationship can be further seen in Fig. 8 where customer main type 1 represents customer sub-type 1-5, customer main type 2 represents customer sub-type 6-8 and so on.
(3) It can be observed from Fig. 6 that the highest number of insurance families are purchased by customer main-type 8 which represents 'family with grownups' whereas customer main-type 4(career loners) has 0 policies. We can also observe that customer sub-type 8 (middle class families) have highest number of insurance policies followed by sub-type 33(lower class families), whereas sub-types 16-19(youth, students) doesn't have nay policies.
(4) It can be observed from Fig. 3 that all insurance policy owners either own 1 or 2 houses. Most of them are in the age group of 40-50 and most common household size is 3.
(5) We can observe from correlation matrix that single people are more likely to own no car, the correlation between 2 columns is 0.51, in comparison correlation between single people owning 1 car is -0.3 and that of 2 cars is -0.2. We can observe that in Fig. 7 that there is a very high percentage of single people having no insurance policies(belonging to 1st bin). We can also observe that married people are more likely to have insurance policies.
(6) From fig. 5 it can be observed that most of the people who own insurance policies are in the income range of 30,000-75,000.
(7) From fig. 4 it can be observed that people with medium or low level education are more likely to purchase insurance policy.

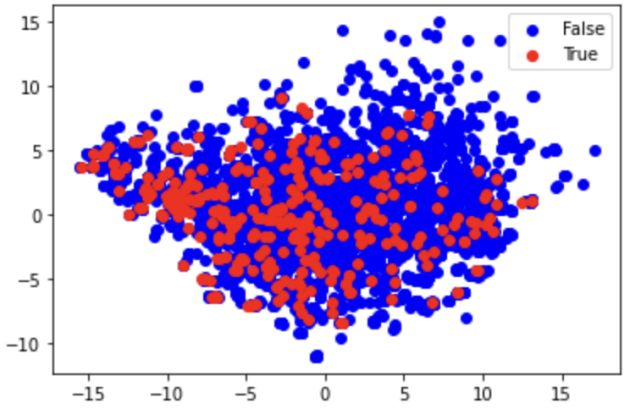
2.3 Dimensionality Reduction

We will use both PCA and t-SNE for dimensionality reduction, the plots are shown in Fig. 1. Here 'True' denotes people who bougth the insurance policy and 'False' denotes the people who did not buy the insurance policy. It can be observed from the Fig. 1 that there is a high degree of overlap between positive and negative classes in 2-dimensions.
It can be observed from Fig. 2 that 2 dimensions explain only 30% of variance in data and 69 dimensions explain 99% of variance in data. This is the first time adding additional dimensional doesn't show any change to the third decimal of variance.

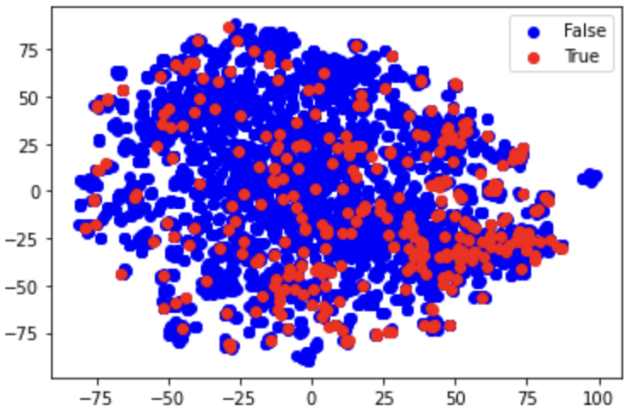
3 METHOD

3.1 Evaluation Metric

As we are trying to evaluate ability of our binary classifier we will be using the receiver operating characteristic (ROC) curve and corresponding area under the curve (AUC). As no test set or validation set is provided, we will perform 10-fold cross validation for each method and report the maximum and mean value of ROC AUC score.



(a) PCA plot



(b) t-SNE plot

Figure 1: Dimensionality Reduction use PCA and t-SNE

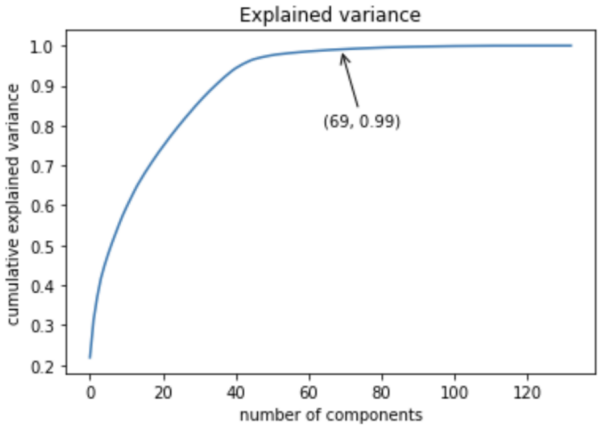
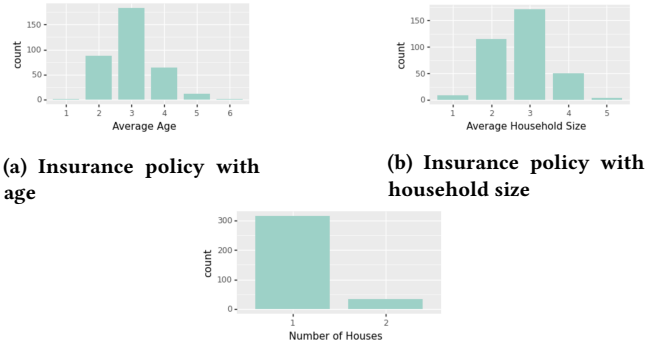


Figure 2: PCA cumulative variance

3.2 Linear Regression

We will use multiple linear regression to model the linear relationship between the independent and dependent variables. This

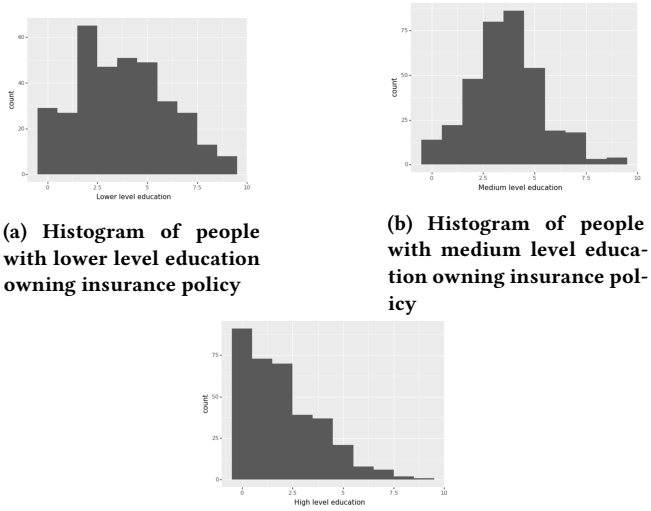


(a) Insurance policy with age

(b) Insurance policy with household size

(c) Insurance policy with number of houses

Figure 3: Relationship of insurance policy with various factors



(a) Histogram of people with lower level education owning insurance policy

(b) Histogram of people with medium level education owning insurance policy

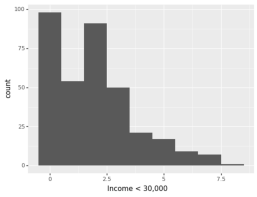
(c) Histogram of people with higher level education owning insurance policy

Figure 4: Relationship of insurance policy with various factors

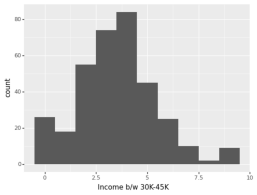
relationship can be represented by the following equation:
$y_i = b_0 + b_1 * x_{i1} + b_2 * x_{i2} + ... + b_p * x_{ip} + \epsilon$
where for $i = n$ observations:

(1) y_i is dependent variable
(2) x_i is independent variable
(3) b_0 is y-intercept or bias
(4) b_p is slope coefficients for each independent variable
(5) ϵ error term

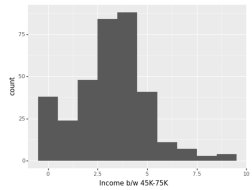
As we are approaching the problem of binary classification using linear regression, during prediction stage we will restrict the output of classifier to range $[0, 1]$ by clipping the values.



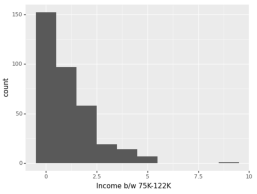
(a) Histogram of people owning insurance policy with income under 30,000



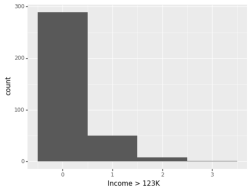
(b) Histogram of people owning insurance policy with income between 30,000-45,000



(c) Histogram of people owning insurance policy with income between 45,000-75,000

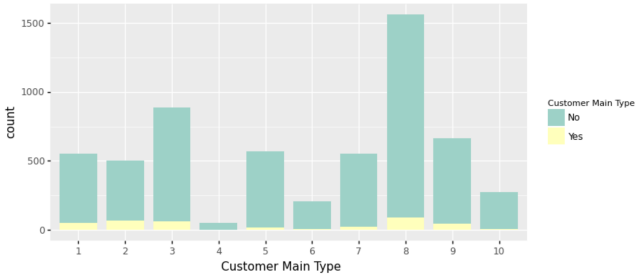


(d) Histogram of people owning insurance policy with income between 75,000-122,000

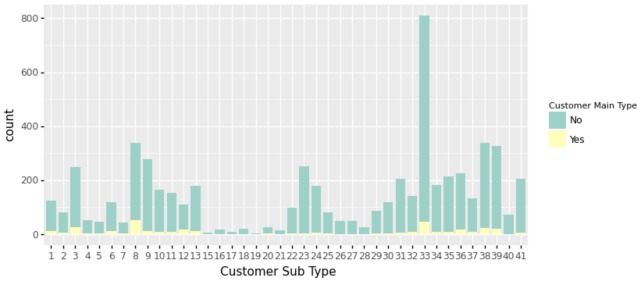


(e) Histogram of people owning insurance policy with income over 123,000

Figure 5



(a) Number of people from customer main type who purchase insurance policy



(b) Number of people from customer sub type who purchase insurance policy

Figure 6

	Original Dataset	SMOTE
Linear Regression	0.79	0.798
Logistic Regression	0.51	0.753

Table 1: Max ROC AUC Score on the test split after 10-fold cross validation

	Original Dataset	SMOTE
Linear Regression	0.73	0.72
Logistic Regression	0.50	0.658

Table 2: Mean ROC AUC Score on the test split after 10-fold cross-validation

3.3 Logistic Regression

We will use logistic regression to model the relationship of dependent variable with independent variable using logistic function. It is generally used when dependent variable is dichotomous. We can represent this relationship through the following equation
$f(x) = b_0 + b_1 * x_{i1} + b_2 * x_{i2} + ... + b_p * x_{ip}$
where for $i = n$ observations:

(1) $f(x)$ is dependent variable or logit
(2) x_i is independent variable
(3) b_0 is y-intercept or bias
(4) b_p is slope coefficients for each independent variable

$p(x) = \frac{1}{1 + e^{-f(x)}}$ where p(x) represents the sigmoid function and is generally interpreted as predicted probability $p(x)$ is close to 1 or not.

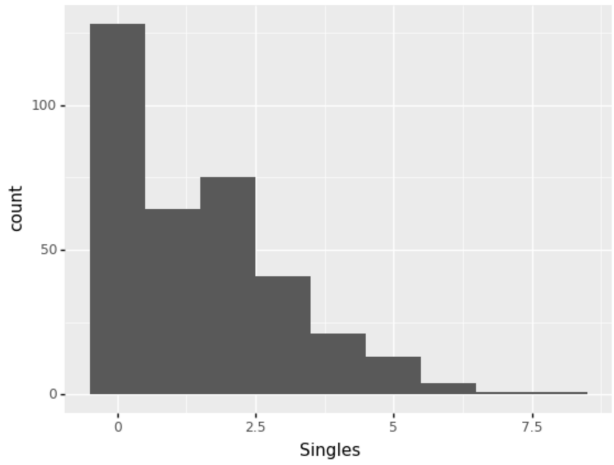
To get the best weights we maximize the log-likelihood function using maximum-likelihood estimation using the equation:
$LLF = \sum_i (y_i * \log p(x_i) + (1 - y_i) * \log(1 - p(x_i)))$

4 OBSERVATIONS

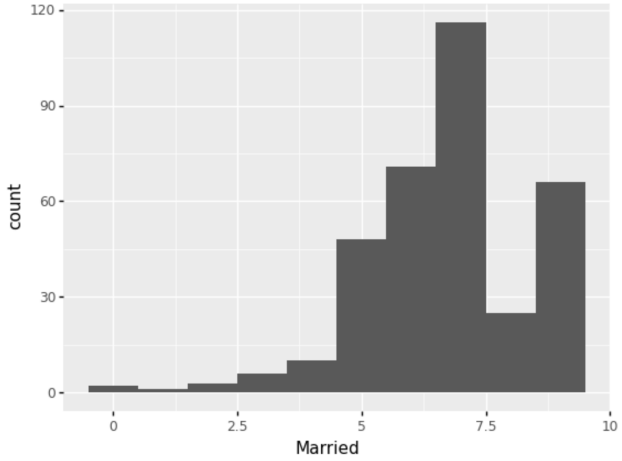
It can be observed that SMOTE based oversampling on average leads to 1.3% decrease in linear regression scores, but 31.6% increase in logistic regression on the test split after 10-fold cross validation.

5 IMPLEMENTATION DETAILS

SMOTE based oversampling was performed after K-Fold cross validation splits, to prevent same samples for being present in both train and test set. For all experiments same 10-split of dataset was used. In logistic regression L2 norm was used with C value(inverse of regularization strength) equal to 1. Categorical columns were one hot encoded and numerical columns were standardized before training as well as before dimensionality reduction. No class weights were used to perform the experiments.



(a) Histogram of single people owning car insurance



(b) Histogram of married people owning car insurance

Figure 7

ROC curves and confusion matrix can be found in the jupyter notebook sent along this report.

6 CONCLUSION

It can be concluded that people belonging to age group of 45-50 years, married, having 1-2 houses, having income between $30,000-$75,000, with average household size 3, and received medium or low level education are more likely to buy caravan insurance policies. We can also observe that linear regression performs nearly similar on original as well as over-sampled dataset but logistic regression shows a jump of 31.6% in performance on over-sampled dataset.

ACKNOWLEDGMENTS

This project was done as part of the MA616: Elements of Data Science course at the Indian Institute of Technology, Ropar, Punjab, India. I am thankful to Dr. Arun Kumar(Department of Mathematics, IIT Ropar) for offering this course and providing all the guidance needed for completion of this project.

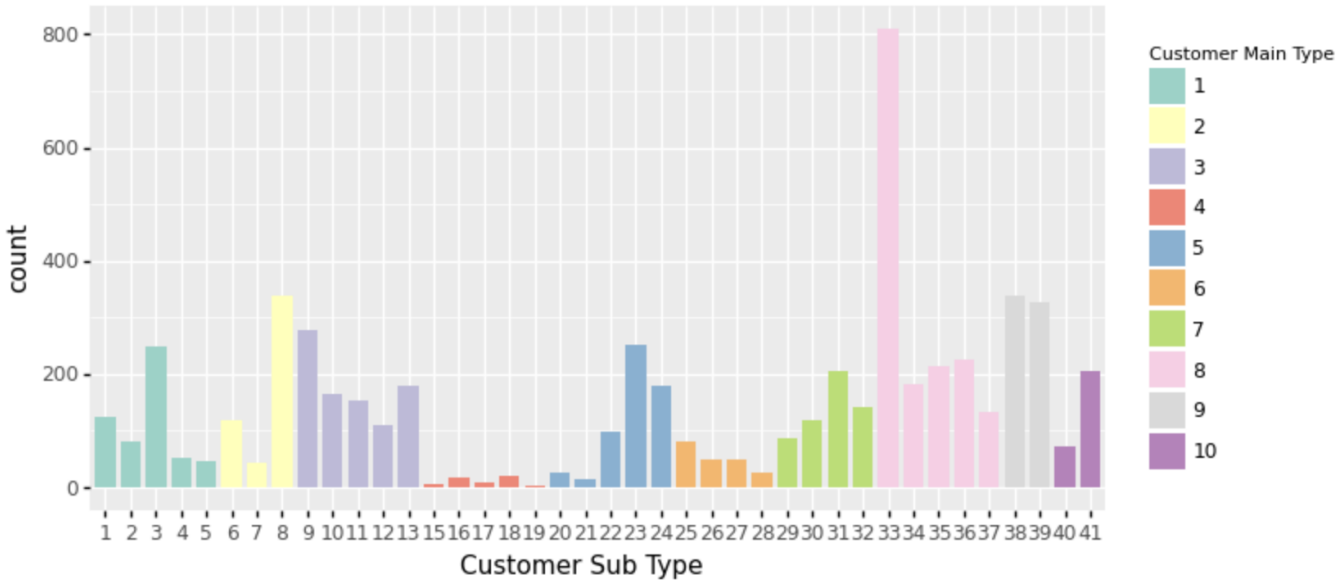


Figure 8: Relationship between customer main type and customer subtype

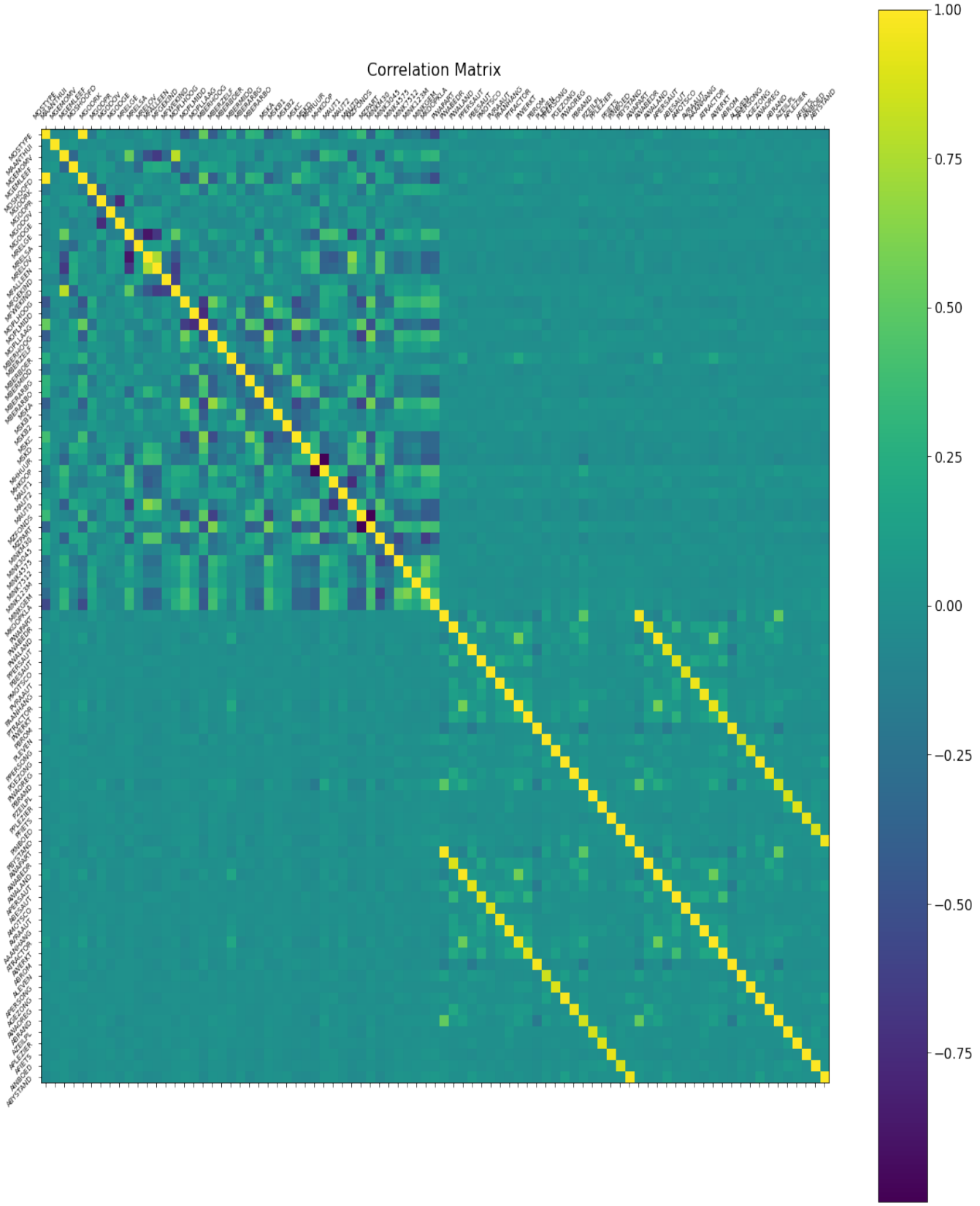


Figure 9: Correlation between various parameters of the dataset

REFERENCES

[1] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority over-Sampling Technique. *J. Artif. Int. Res.* 16, 1 (June 2002), 321–357.