ZS Case Study for Young Data Scientist- Oct '17

RARE DISEASE PREDICTION

Submitted By -
Subham Agrawal
IIT Madras

Quality Checks

- No missing data
- No duplicates
- Data types:
  - **Continuous**: var5, var9, var15, var17, var18, var19
  - **Binary**: var6, var7, var8, var13, var23, var24, var25, var26
  - **Categorical/Ordinal**: var1, var2, var3, var4, var10, var11, var12, var14, var16, var20, var21, var22
  - Categorical/Ordinal variables have few categories which might be encodings of different symptoms of diseases. It is difficult to say if these attributes have some intrinsic ordering.

Quality Checks(Contd…)

- Training dataset is **HIGHLY IMBALANCED** (19400:600) with ratio(>30).
- From correlation matrix of features,

  (var5, var9) & (var17, var18) has high correlation (i.e. >0.5)
- Inconsistency:

  Categories/range of var1, var20 and var22 are different in training and testing dataset.
- Resident ID's are not same in training and test data. This suggests that algorithms like recommendation system are not a good choice.

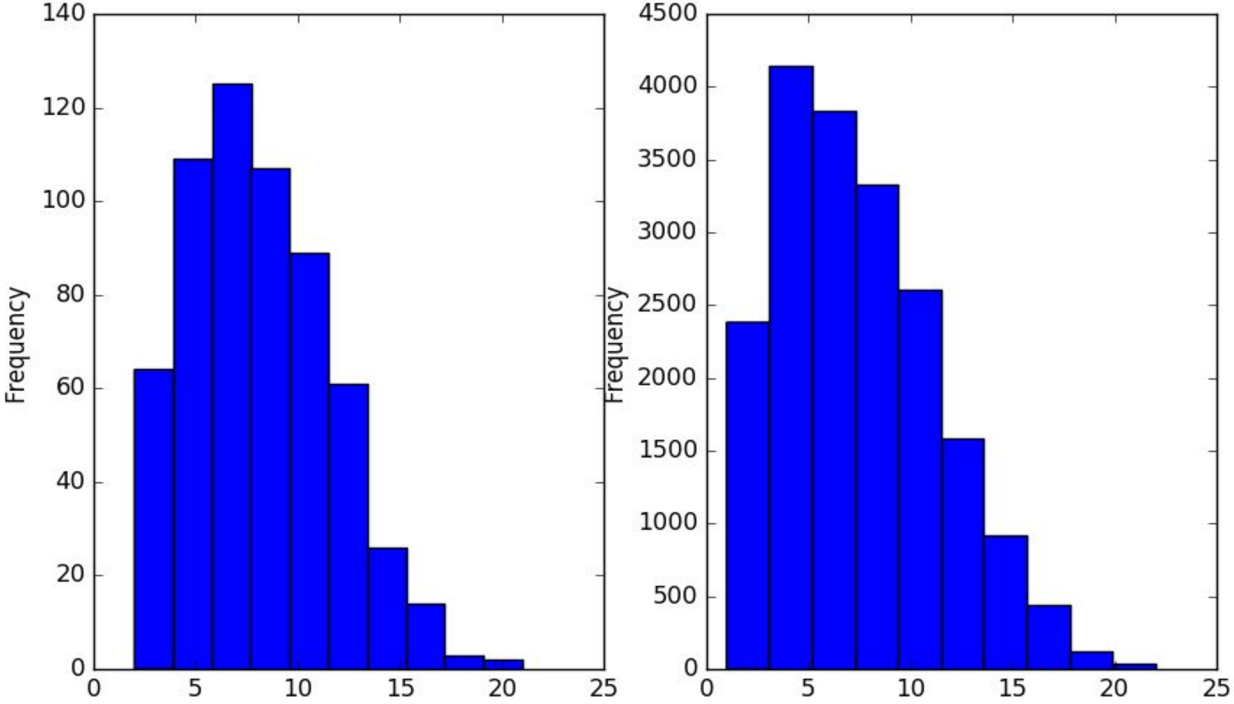
Data Preprocessing

- **Sampling techniques**:
    a. **Undersampling:** A dataset with undersampling of Disease Flag 0 is generated.
    b. **Oversampling (SMOTE):** A dataset with oversampling of Disease Flag 1 is generated such that dependent variable is balanced.
- **Feature generation:**
    a. Dummy features are generated for variables which may be either ordinal/nominal.
    b. Continuous and binary features are not changed.

Key Observations/Trends

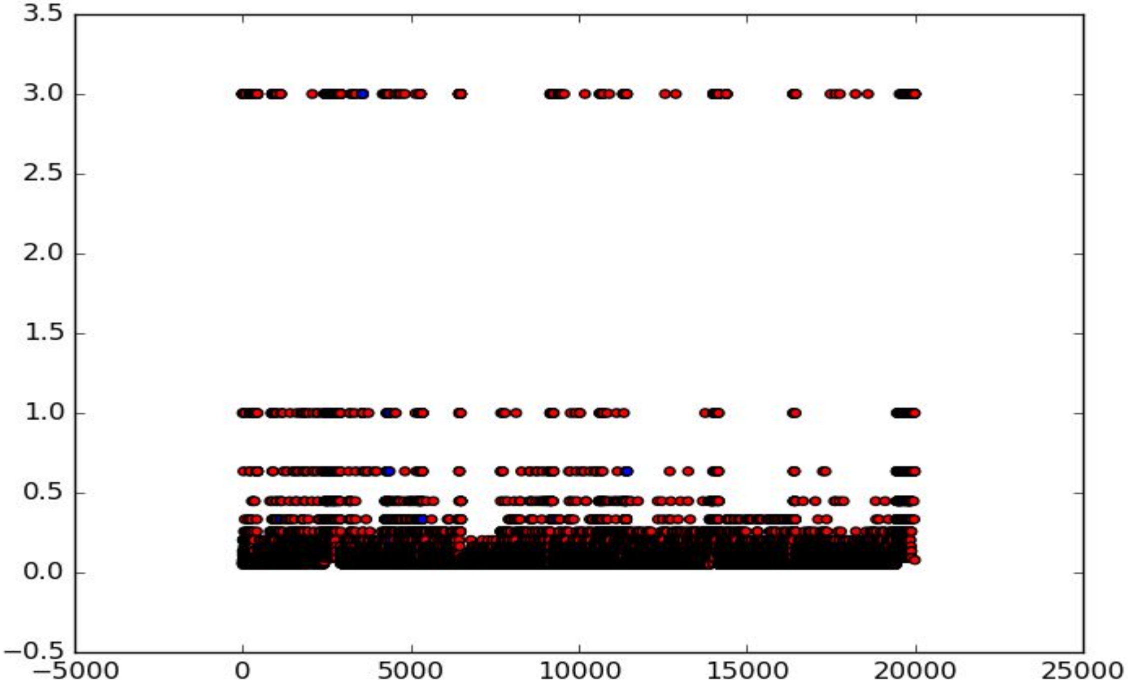
Plotted frequency histograms of each variable for disease flags 0 and 1. For all variables in training dataset, both frequency plots were similar which didn't indicate any trend.

Example,



Key Observations/Trends (Contd…)

No pattern was observed in scatter plots for each variable. In scatter plots, reds points denote ‘no disease’ class and blue points denote ‘disease’ class. Example,



Scatter plot

Red color-No disease

Blue color- Disease predicted

Model Choice Explanation

XGBoost

- Since the data is imbalanced and no trend is observed, bagging models were a good start.
- Later, XGBoost was implemented to achieve higher accuracy because:
    - It takes care of class imbalance with no preprocessing/sampling techniques.
    - Another benefit is that they can automatically provide estimates of feature importance from a trained predictive model and there is no need to do feature selection.
    - For XGB model, all the variables as well as dummy of ordinal/nominal variables are taken as features. This can be done because adding redundant features does not affect XGB accuracy significantly.

Model Choice Explanation (Contd…)

XGBoost

- Tuned XGBoost to achieve a mean AUC score of 0.57 with cross validation=10 and the same was achieved in the test data.
- Tuned parameters are as follows:
  - learning rate = 0.05
  - number of estimators = 100(default)
  - objective - ’binary:logistic’
  - scale_pos_weight = 1.5
  - Setting weight for classes to increase the cost of error for minority class increased accuracy.

Model Choice Explanation(Contd…)

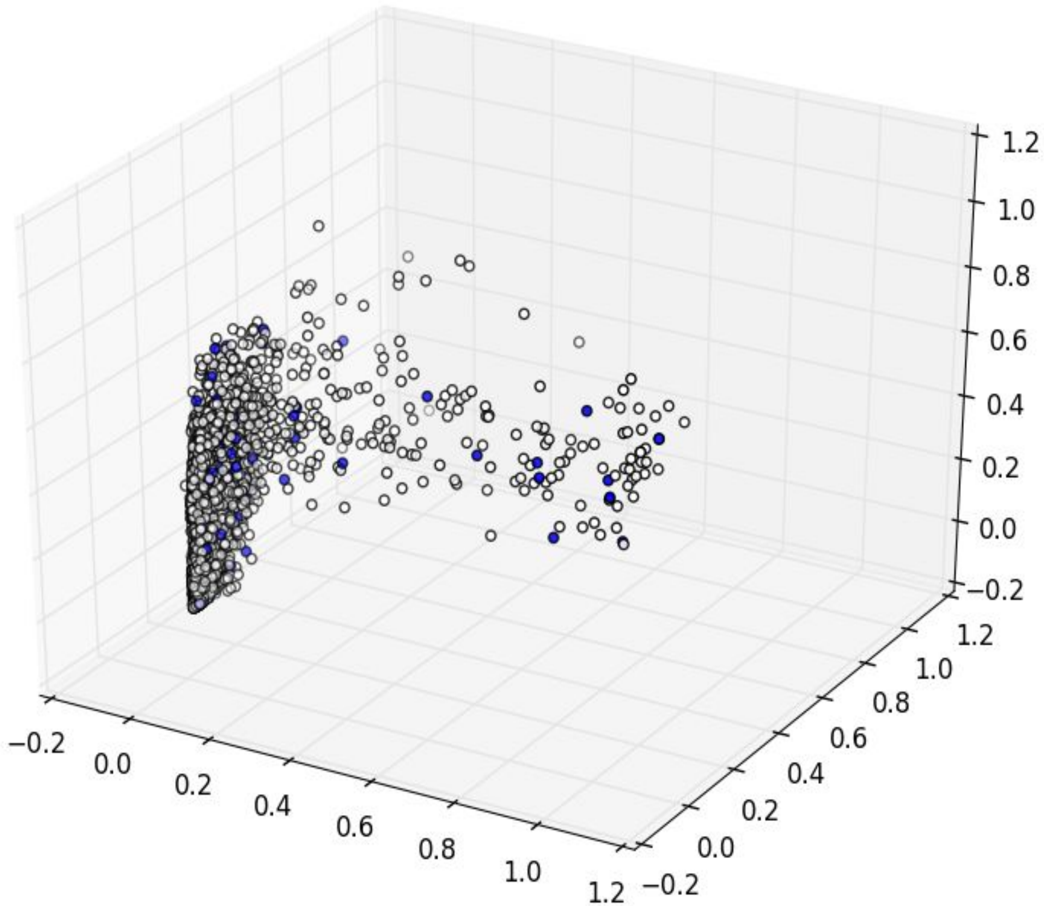
Naïve Bayesian

- Next best choice was Naïve Bayes classifier because variables in dataset can be symptoms which can cause disease and it's popularly used for disease prediction.
- To implement Naïve Bayes classifier in a dataset with mixture of categorical and numerical variables two models were trained.
  1. **Gaussian Naive Bayes** - continuous variables as features
  2. **Multinomial Naive Bayes** - binary variables and dummy of ordinal/nominal variables as features.
- Average of probabilities from both models gave an **AUC score of 0.58** in the test data.

Model Choice Explanation(Contd…)

- Both models (i.e. XGBoost and Naïve Bayes) were performing equally good with an AUC score of **~0.6**
- Sampling techniques were not effective.
  a. **Undersampling**: Reduced AUC score because useful data points were removed.
  b. **Oversampling(SMOTE)**: It led to over-fitting of decision-based learning methods and decrease in AUC score of Naive Bayes classifier.
- Probabilities from all the models were normalized and then equal weights were given to each model. It increased AUC score to 0.605.

Model Choice Explanation (Contd…)



Plotted probabilities from all the 3 models, so as to be able to apply classification/clustering techniques. Unfortunately no pattern was observed.

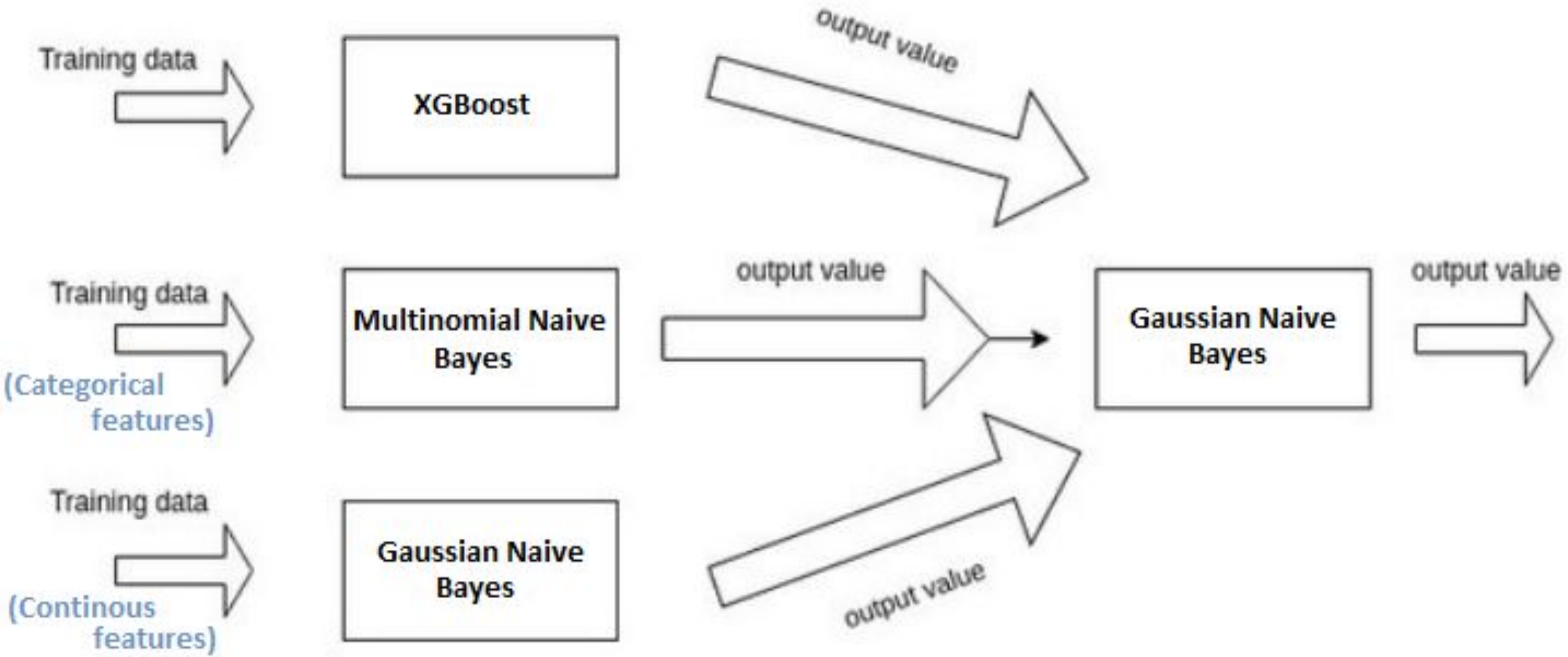
In this 3-D plot, blue dot represents 'disease' class and white dot represent 'no disease' class.

Model Choice Explanation (Contd…)

ENSEMBLE

- Ensemble techniques is used to modify existing classification algorithms to make them appropriate for imbalanced dataset.
- Thus, main objective of ensemble methodology is to improve the performance of single classifiers i.e. XGBoost and Naïve Bayesian.
- **Gaussian Naïve Bayesian** classifier was finally chosen for ensemble of all three different models.
- Probabilities from each model was normalized for input to ensemble classifier.
- This improved the performance of the model to achieve an **AUC of 0.6134** which is the final submission.

Model


Training data
XGBoost
output value
Training data
Multinomial Naive Bayes
(Categorical features)
output value
Gaussian Naive Bayes
output value
Training data
Gaussian Naive Bayes
(Continous features)
output value

Features

- **XGBoost**
  - A benefit of using ensembles of decision tree methods like gradient boosting is that they can automatically provide estimates of feature importance from a trained predictive model and there is no need to do feature selection.
  - Thus, all the variables (1 to 26) and dummies of ordinal/nominal variables are taken as input features.
  - No other preprocessing is done for XGBoost.
- **Gaussian Naïve Bayesian** - All the continuous features are taken as input for this Naïve Bayesian classifier.
- **Multinomial Naïve Bayesian** - All binary and dummy of ordinal/nominal variables are considered as features.

Expected Error of Submission

- Method used to cross-validate using train data are as follows:
  - Split training dataset in 80-20
  - Cross-validate single classifiers(XGB and Naive Bayes) with cv=10 using 80% of training data and tune models.
  - Train each classifier using same 80% of training data
  - Use trained models to make prediction on rest 30% training data
  - Cross-validate ensemble Naïve Gaussian model with inputs as normalized predicted probabilities (30% of training data).
- **Mean AUC** score is **0.623** with **standard deviation 0.015** of this final cross validation of ensemble model.
- Thus, **no over-fitting** in the final submission.

Feature Importance

- Most significant features with their predictive power as obtained from XGBoost trained model using feature importance method are:
  1. Var18 - 0.5
  2. Var3 - 0.22
  3. Var4 - 0.10
  4. Var9 - 0.09
  5. Var11 - 0.04

Thank You