Unit 11—Risky Business

Background

Auto loans, mortgages, student loans, debt consolidation ... these are just a few examples of credit and loans that people are seeking online. Peer-to-peer lending services such as LendingClub or Prosper allow investors to loan other people money without the use of a bank. However, investors always want to mitigate risk, so you have been asked by a client to help them use machine learning techniques to predict credit risk.

In this assignment, you will build and evaluate several machine-learning models to predict credit risk using free data from LendingClub. Credit risk is an inherently imbalanced classification problem (the number of good loans is much larger than the

number of at-risk loans), so you will need to employ different techniques for training and evaluating models with imbalanced classes. You will use the imbalanced-learn and Scikit-learn libraries to build and evaluate models using the two following techniques:

1. Resampling
2. Ensemble Learning

Files

Resampling Starter Notebook

Ensemble Starter Notebook

Lending Club Loans Data

Instructions

Resampling

You will use the imbalanced learn library to resample the LendingClub data and build and evaluate logistic regression classifiers using the resampled data.

You will:

1. Load the Lending Club data, split the data into training and testing sets, and scale the features data.
2. Oversample the data using the `Naive Random Oversampler` and `SMOTE` algorithms.
3. Undersample the data using the `Cluster Centroids` algorithm.
4. Over- and under-sample using a combination `SMOTEENN` algorithm.

For each of the above, you will need to:

1. Train a `logistic regression classifier` from `sklearn.linear_model` using the resampled data.
2. Calculate the `balanced accuracy score` from `sklearn.metrics`.
3. Calculate the `confusion matrix` from `sklearn.metrics`.
4. Print the `imbalanced classification report` from `imblearn.metrics`.

Use the above to answer the following:

> Which model had the best balanced accuracy score?
>
> Which model had the best recall score?
>
> Which model had the best geometric mean score?

Ensemble Learning

In this section, you will train and compare two different ensemble classifiers to predict loan risk and evaluate each model. You will use the Balanced Random Forest Classifier and the Easy Ensemble Classifier. Refer to the documentation for each of these to read about the models and see examples of the code.

Be sure to complete the following steps for each model:

1. Load the Lending Club data, split the data into training and testing sets, and scale the features data.
2. Train the model using the quarterly data from LendingClub provided in the `Resource` folder.
3. Calculate the balanced accuracy score from `sklearn.metrics`.
4. Print the confusion matrix from `sklearn.metrics`.
5. Generate a classification report using the `imbalanced_classification_report` from imbalanced learn.
6. For the balanced random forest classifier only, print the feature importance sorted in descending order (most important feature to least important) along with the feature score.

Use the above to answer the following:

> Which model had the best balanced accuracy score?
>
> Which model had the best recall score?
>
> Which model had the best geometric mean score?
>
> What are the top three features?

Hints and Considerations

Use the quarterly data from the LendingClub data that is provided in the `Resources` folder. Keep the file in the zipped format and use the starter code to read the file.

Refer to the imbalanced-learn and scikit-learn official documentation for help with training the models. Remember that these models all use the model->fit->predict API.

For the ensemble learners, use 100 estimators for both models.

Submission

- Create Jupyter notebooks for the homework and host the notebooks on GitHub.
- Include a markdown that summarizes your homework and include this report in your GitHub repository.
- Submit the link to your GitHub project to Bootcamp Spot.

© 2020 Trilogy Education Services, a 2U, Inc. brand. All Rights Reserved.