Project Writeup - Risk Free Invesment (Skewed Data)

Objectives

Below is the list of questions that we are looking to answer as a final outcome of this project:

- To generate featured data and provide prediction whether a particular loan will be fully paid off on time or not based on developing a Predictive Modeling using **Skewed/Imbalanced Credit Risk data** set using a Supervised Machine Learning Algorithm.

Goal Significance

Why investors need to know the certainty of getting loan pay-off? What benefit we could derive by developing accurate loan delinquency prediction model?

We can highlight below mentioned goals:

1. The investors, especially conservative investors, tend to stay away from any risky investment opportunity where the probability of default or charge off is high. They want assurance of loan pay off on time and per T&C agreed.
2. Investors also desire not to miss any promising business opportunity due to faulty predictive model, where it is highly likely that the load would get pay off on time.

Data

Data Source

Lending Club periodically releases data for all of the approved and declined loan applications periodically on their website (www.lendingclub.com) for the use of their investors to make loan offer decisions. The investor can select a few different year ranges to download the datasets (in CSV format) for both approved and declined loans for his/her use.

The data file download is available from: Google Docs

Data Lists

The data set files:

1. Particulars of the features: `LCDataDictionary.xlsx`
2. Potential clients' data: `loans_2007.csv`

Data Extraction Details

The data set contains total 52 features for 42535 candidates with the particulars of these feature is listed in the `LoanStats` tab of `LCDataDictionary.xlsx` file.

The following features are included in the data set:

Project Writeup - Risk Free Invesment (Skewed Data)

Id, member_id, loan_amnt, funded_amnt, funded_amnt_inv, term, int_rate, installment, grade, sub_grade, emp_title, emp_length, home_ownership, annual_inc, verification_status, issue_d, loan_status, pymnt_plan, purpose, title, zip_code, addr_state, dti, delinq_2yrs, earliest_cr_line, inq_last_6mths, open_acc, pub_rec, revol_bal, revol_util, total_acc, initial_list_status, out_prncp, out_prncp_inv, total_pymnt, total_pymnt_inv , total_rec_prncp, total_rec_int, total_rec_late_fee, recoveries, collection_recovery_fee, last_pymnt_d, last_pymnt_amnt, last_credit_pull_d, collections_12_mths_ex_med, policy_code, application_type, acc_now_delinq, chargeoff_within_12_mths, delinq_amnt., pub_rec_bankruptcies, tax_liens.

Model

How was data cleaning achieved?

As we read the data set available from the Lending Club, the columns with the following characteristics were eliminated:

- that had data duplication issues,
- that contained superfluous/irrelevant information,
- that carry single value only,
- that carry more than 1% null values
- that needed additional analysis to turn into useful features.

The features columns that had formatting issues were also cleaned and converted in to categorical columns with dummy variables.

How were features prepared?

- Manipulate the available features and introduced some new features
- Select target column and clean it to narrow down predictive features of interest
- Extract the data pertaining to the new features
- Clean the data to facilitate the new features adjustments
- Map certain columns with multiple categorical data
- Convert string type to numeric type data to facilitate further analysis
- Verify and confirm data skewness
- Setting up hypothesis for the model prediction.

Why only this model?

For classification predictive need, Logistic Regression provides one of the suitable supervised learning algorithms for the given type of data set. Logistic Regression gives better model to give rise reliable generalized predictive capability over the cross-validation set and other external data.

Project Writeup - Risk Free Invesment (Skewed Data)

Highlights of the code.

Software packages used:

- Python
- Pandas
- Numpy
- Seaborn
- Matplotlib.pyplot
- Sklearn-

  LogisticRegression, cross_val_predict, mean_square_error, classification_report, confusion_matrix, GridSearchCV
- Mlxtend (plot_confusoin_matrix)

Overview:

- Read the data and form the dataframe
- Initialize new features
- Generating moving average and variance for different durations
- Update data frame with the data for the new features
- Data Cleaning, data slicing and development of training/test/CV sets
- Feature identification and categorical features development
- Development of predictive model using logistic Regression
- Feature significance assessment and model fit evaluation
- Logistic algorithm predictive verification (for model & revised model)
  - Fitting the model over the test set
  - Verifying the model predictability over the cross-validation set
  - Confirming the model generalization using another classifier (*K-Fold Cross Validation* & *GridSearchCV*)
- Check for model predictive output confidence to ascertain stringent investors:
  - Absolute removal of 'False Positive'
  - Strong commitment for 'True Positive'

How does the data fit to the model?

- The model training over the test set was assessed by checking the MSE value of the error matrix.

Model Validation Details.

- Model was evaluated revised once to ensure:
  - it has learned the imbalance data correctly, and
  - it takes into account the data skewness appropriately for unbiased predictability
- Logistic algorithm predictive verification (for model & revised model)
  - Fitting the model over the test set
  - Verifying the model predictability over the cross-validation set

  - Confirming the model generalization using another classifier over the cross validation set (using *K-Fold Cross Validation* & *GridSearchCV*)
- Check for model predictive output confidence to ascertain stringent investors:
  - Absolute removal of 'False Positive'
  - Strong commitment for 'True Positive'

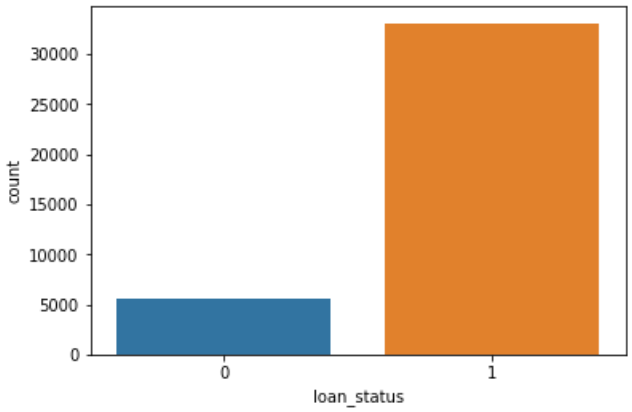
Justification for the meaningfulness of the model outcome.

- For the imbalanced/skewed data, the model predictability results show zero error (MSE) over the test set and cross validation set.
- The model prediction results show precision = 1.0, recall = 1.0 and FPR = 0.0 over cross validation set. This indicates 100% reliability for predicting results.
- For more details, please see Conclusion section of the Jupyter Notebook.

Results

Visualize the results.

- **Data Skewness Matrix**

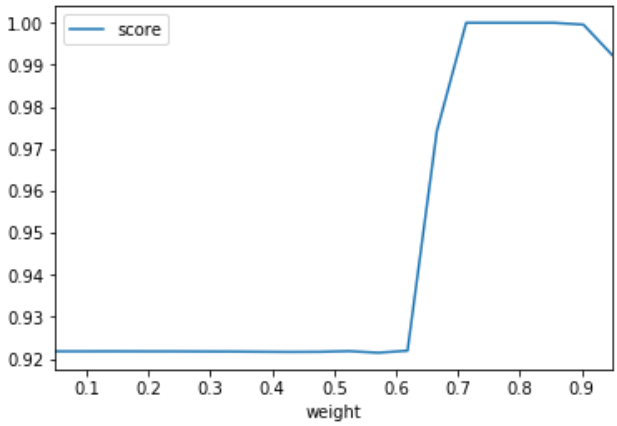


- **Model Generalization and Predictivity Check:**

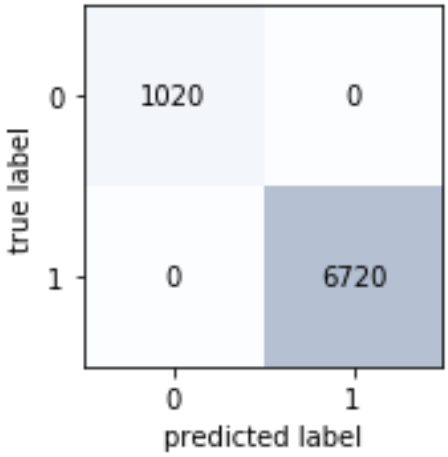
	Test Set	Cross Validation Set	GridSearch CV
precision	1.0	1.0	0.999703
recall	1.0	1.0	1.000000
accuracy	1.0	1.0	0.999742
fpr	0.0	0.0	0.001961
tpr	1.0	1.0	1.000000
F1_score	1.0	1.0	0.999851
MSE	0.0	0.0	0.000258

- **GridSearchCV Best Parameters Class Weight Distribution**

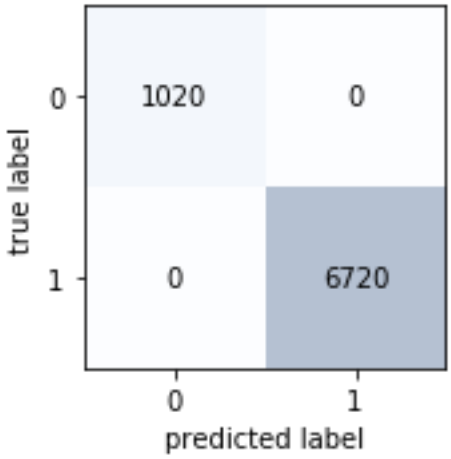
Project Writeup - Risk Free Invesment (Skewed Data)



- **Confusion Matrix using GridSearchCV Classifier over Cross-Validation Set**

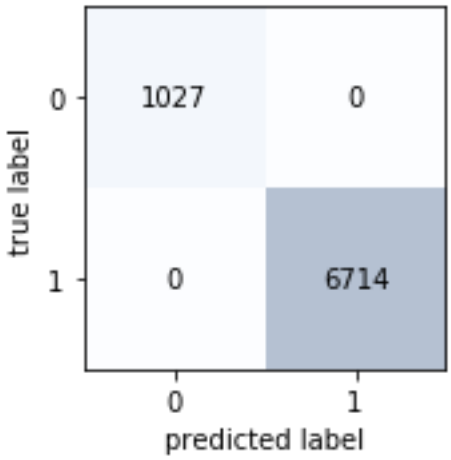


- **Confusion Matrix for Logistic Reg Prediction over the Cross-Validation Set**

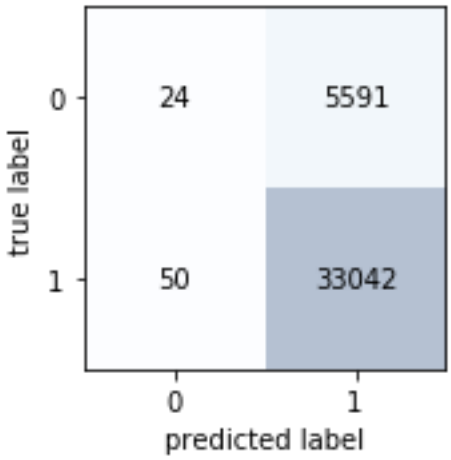


Project Writeup - Risk Free Invesment (Skewed Data)

- **Confusion Matrix for Logistic Reg Training over the Test Set**



- **Confusion Matrix for K-fold Cross-Validation over the Entire Data Set**



Explain the results in a simple, concise and easy way. (non-technical audience)

- Model prediction performance is unbiased for the imbalance credit risk data set.
- The Model's overall prediction reliability for the other external data base is very high.
- The model recommends investment recommendation ONLY to really safe loan options that are going to be paid off on time. In other words, the model never suggests any potentially risky investment opportunity as a safe option.
- The model prediction never misses any strong investment opportunity by wrongly identifying a safe borrower as a potential delinquent, i.e. loan that would NOT be paid off on time. In other words, if model drops any opportunity then it is highly likely that the option would be very risky.
- The model suggestions would be extremely helpful for conservative investors to make appropriate investment decisions with greater confidence.

Project Writeup - Risk Free Invesment (Skewed Data)

Explain the results in the most technical way. (technical, data-scientist audience)

1. The results above show generalization of the revised model over the unknown data set.
2. Precision value of 1.0 indicates that this model predicts ONLY actual positive examples as positive outcome.
3. Recall value of 1.0 is an indication that this model captures ALL positive examples correctly.
4. F1-Score is an indicative of perfect 'Precision' and 'Recall' scores.
5. The zero value of fpr confirms that the model is absolutely not predicting "False Positive". This confirm the reliability of the algorithm and its alignment with our need for conservative business proposition to avoid any risky investment option. The checks for model predictive behavior above assure this model would recommend only loan options that are safe and will be paid off on time.
6. The TPR value of 1.0 is an indicative that this model will show ALL of the safe loan opportunities that will be paid off on time. This will give an utmost confidence to the conservative investors and encourage them to treat any/all of the recommended business opportunity as risk free to do business.

Conclusion

What we learn from this outcome. (non-technical audience)

- The results give rise encouragement to the existing and potential conservative investors to rely with confidence on the model investment suggestions.
- The model performance reliability, in turn, discourage the loan seekers with poor credit history to take advantage of any system loop holes.
- The model would give very clear message to all credit seekers how credit merit is rewarded. This would encourage the creditors to be responsible, rectify their past mistakes and improve their credit worthiness.

Technical significance of the results. (technical, data-science audience)

- The model predictions are verified not only on the test set, but also over the cross-validation set. This enhances the confidence for the model generalization.
- The model behavior is evaluated for the imbalanced credit data and its unbiased classification is confirmed by the external class weight GridSearchCV classifier over the cross-validation set.
- F1-Score of 1 and FPR value of 0 confirms the no “False Positive” and minimal “False Negative” predictive behavior of the model.

Project Writeup - Risk Free Invesment (Skewed Data)

Suggestion for Further Development

How differently you would have done this project, if get a chance to do it again. Why?

1. I could have involved some other algorithm such as Random Forest and ensemble multiple models to assess the predictive model reliability over the skewed data.

2. It would also be very interesting to repeat and recheck the model behavior by putting more data from the same source and/or using the similar data from some other comparable market player.

Your suggestions for someone if s/he wants to further continue this work.

Someone could:

- pick/drop one or more features that are already used/not used in this model
- try models using classifier other than Logistic regression and GridSearchCV
- tweak further the class weightages for binary predictions

to enhance the training algorithm and achieve higher performance.