Contents

1 Literature Review

1.1 Overview of data imbalance problem

Classification is the most common machine learning tasks, giving practical solutions to variety of field including financial analytics. When dealing with binary classifications, in many use cases the interest class - the class that holds value in predicting outcome - is extremely smaller in size. In theory, ML models tend to favour the majority class, leading to biasness. Such problem is addressed as a data imbalance issue. Data imbalance is a well-recognised issue in building such ML classification models. According to Haixiang, G. et al. (2017), about 43.5% of the research publications in machine learning limitations are associated with data imbalance in the past decade in financial services. As highlighted by Haixiang, G. et al. (2017), there are many publications that propose techniques to handle the class imbalance issue. These are broadly classified into: data-level approach and algorithmic level approach:

1.1.1 At the Data Level

Sampling techniques modifies the class distribution in the data set before model training is conducted. It recalibrates the class skewness to decrease predictive bias due to majority class. There are three major techniques:

1. Oversampling such as random oversampling, SMOTE (synthetic minority oversampling technique) (Chawla et al., 2002).

2. Under sampling such as random under sampling (Estabrooks et al., 2004)
3. Hybrid sampling which are mix of the above two methods. SMOTEENN (Batista et al., 2004), a combination of SMOTE and Edited Nearest Neighbours is a well-known hybrid approach.

1.1.2 At Algorithm Level

In this method, the model learning is modified during training to handle data imbalances.

1. Ensemble methods: algorithms like AdaBoost (Freund and Schapire, 1997) and gradient boosting (Friedman, 2001), balanced bagging (He and Garcia, 2009).
2. Model classifiers: Models that have inherent function to adapting data imbalances. These models such as Support Vector Machines (Cortes C. & Vapnik V., 1995), Decision Trees (Quinlan J., 1986), Random Forest (Breiman L., 2001), etc. are tuned to allow class weights in giving a high priority or weight to minority class. Hence, the models are less prone to overfitting and biases.

1.1.3 Cost Sensitive learning (CSL) framework

The cost sensitive models can be applied at both data and algorithm level (Elkan, 2001).

1. At Data Level: Instance weighing (Ting K.M., 1998), assigns misclassification costs during the training phase by adding different weights to pay more attention to minority class.
2. At Algorithm Level: MetaCost (Domingos P.,1999) is a CSL technique where the model is transformed to cost sensitivity during model learning. In this technique, the algorithm is re-adjusted to different thresholds according to the associated misclassification costs.

The current literature for each of the three techniques is discussed below.

1.2 Data level (Pre-Processing techniques)

The data level sampling approach is the most popular technique to handle data imbalance issue (Susan & Kumar, 2020). This process happens at the data level before feeding into the machine learning models for prediction. In order to achieve a certain equilibrium between the two classes before training into a ML model, it includes adjusting the distribution of majority and minority class by either done in two ways:

- Increasing number of samples of minority class to match majority class, known as ‘over-sampling’.
- Decreasing the majority class to match minority population, called ‘under-sampling’.

- Combination of the above two is known as mixed or hybrid sampling

Although sampling has been found a popular method among researchers in handling class imbalance problem, Susan and Kumar (2020) in their systematic review show these conventional sampling methods on real life dataset applications add up noise and unwanted duplication, which could potentially lead to over/under fitting the data. At the same time, balancing the inter class proportion also makes it challenging to main intra-class diversity (Susan and Kumar, 2020). Another disadvantage could be the loss of useful samples in the sampling process that could have helped in prediction (Fernando et. Al, 2022). Specifically, under-sampling tends to remove essential data points in the learning process, while over-sampling may result in overfitting and sometimes simply multiply the computational cost.

In the domain of probability of default prediction, researchers have employed synthetic minority oversampling technique, known as SMOTE (Chawla et al., 2002), and have yielded superior results on imbalanced datasets. Chawla et al. (2002) uses distance-based K-nearest neighbour methodology to create synthetic samples within the smaller class. The samples that have similar feature characteristics are plotted close to one another.

1.3 Algorithmic Level (Model based methods)

As opposed to data-level approach, algorithm-level methods alter the classification model to handle data imbalance issue. They are broadly categorised into ensemble learning (Liu et al., 2020) and cost sensitive learning (Arya Iranmehr et al., 2019). Most common ensemble learning are: Bagging (Wang et al., 2015) and Boosting (Blagus R. & Lusa L., 2017).

Interaction between sampling techniques and ensemble have been developed to handle the issue at data-level. Some well-known ensemble classifiers that use oversampling are SMOOTE*Boost* (Chawla et al., 2003), 'SMOTE*Bagging'* (Wang and Yao 2009) and RAMO*Boost* (Chen et al. 2010). Some under-sampling ensemble learners are Under*Bagging* (Tao et al. 2006), RUS*Boost* (Seiffert et al. 2010). Combined methods of sampling and boosting models have gained attraction over the decade among researchers. For instance, Maciej Zięba and Tomczak (2014) compares the above-mentioned techniques along with cost sensitive techniques with a Support Vector Machine (SVM) model. Researchers recommend a combination of adaptive oversampling with boosted SVM model for a higher performance in predicting default. Additionally, cost-sensitive pruning of decision tree has been presented as an effective technique in handling class imbalance (Zadrozny B. & Elkan C., 2001). Although, each model can be customised to each use case, it can lead to inductive bias within the model, i.e model makes generalised relationships from the training data. For example, decision tree distributes data into homogenous subsets using hierarchical splits. The tree grows deeper when the data is split into any smaller subsets. When dealing with imbalanced data, such deeper fragmentations may at times prevent tree from learning patterns of the minority class, which is of more value in real life. Such phenomenon is referred to as data fragmentation (Fernández et al., 2017).

Many researchers have handled the issue with ensemble machine learning classifiers. Uddin et al. (2023) compared Logistic Regression, tree based models (such as Decision Tress, Random Forest, Extra Tree), Support Vector Machines (SVM), K-nearest neighbour (KNN), Gaussian Naïve Bayes, and boosting models (such as AdaBoost and Gradient Boosting) to predict loan approval. While other researchers utilized deep learning models (such as neural networks) to predict risk exposure (Wang et al., 2019). With large number of comparative-based research conducted in this field, there a few classifiers found effective in predicting default. These are tree-based models (Wang et al., 2020), SVM (Chen S. et al., 2006) and extreme gradient boosting (XGBoost) (Guo and Zach Zhizhong Zhou, 2022) have been found to outperform in predicting default.

1.4 CSL framework (both at data and algorithm level)

Assuming higher weights or costs for the misclassified or minority class (in this case valuable for prediction), with compared to the majority class (Domingos P.,1999). This is often done by specifying ‘cost matrices’ (Elkan C.,2001). It can both be incorporated at the data level and algorithmic level when building a predictive classification model.

1.4.1 Instance weighing cost-sensitive learning

A variety of approaches within cost-sensitive framework have been put forward to handle the data imbalance issues. Ting (2002) proposes an algorithm by weighing sample size based on the misclassification cost. This cost-sensitive weighing (CSW) method changes the class distribution so that the model is less likely to create a predictive bias. Thus, improving the accuracy of the model when higher misclassification costs are introduced. Weiss (2004) in his review found cost-sensitive learning to be an effective solution to class imbalance problems. Such methods are also applied into different use cases within credit risk. Bahnsen et al. (2015) and Zakaryazad & Duman (2016) experimented CSW framework into fraud detection use cases. Li et al. (2021) incorporated CSW into multiple ensemble models in order to increase predictive accuracy in detecting early defaults. Decision trees with light gradient boosting (ML-Light GBM) was found to be the most effective in such scenario.

1.4.2 MetaCost learning

MetaCost was first introduced by Domingos (1999). The cost modification is done within the model algorithm to assign higher weight to the minority class. The algorithm relies on internal cost-sensitive classifier in order to relabel classes in training sets. MetaCost pays higher attention to the minority class and shown to increase sensitivity of the learning model (Ting, 2000). Such cost sensitive model-based approach has seen multiple use cases within credit risk. Kim et al. (2012) compares cost sensitive two approaches: cost sensitive classifier (Witten & Frank, 2005) and MetaCost (Domingos, 1999) on three credit loan data imbalance datasets and finds MetaCost to be an effective solution in decreasing the classification costs, which in turn increases the predictive power of the model. While Liu et al. (2022) uses MetaCost with boosted tree models to improve predictability in credit scoring, resulting in increasing efficiency and accuracy of assigning credit score.

There have been a number of research in being done to find the ideal solution to class imbalance problem that universally applies to all credit risk models (Jiang et al., 2023). In reality, there is no one definitive answer. On one hand, the higher the imbalance ratio (IR) in the dataset, poorer is the performance of the classification model. Additionally, feature dimensions and dataset characteristics are unique to every use case in credit risk. On the other hand, the efficacy of current techniques is limited and faces drawbacks. For instance, the most extensively used technique, SMOTE, fails to consider data distribution comprehensively. Conversely, cost sensitive learning poses constraints on class weight adjustment, leading to interpretability challenges (Niu et al., 2020). A modern approach to integrate explainable AI has been gaining attraction in the recent years. Chen Y. et al (2024) compares two sampling methods: Local Interpretable Model-agnostic Explanations (LIME) & (SHAP) to tackle class imbalance issue using Extreme Gradient Boosting model.

Each method has its own set of benefits and drawbacks. As compared to resampling methods, cost sensitive learning is more computationally efficient and therefore making it suitable for big data analysis (Khan et al., 2018). However, this method is less popular than other techniques in handling data imbalance issue. According to Krawczyk et al. (2014), it might be due to two reasons: i) It is difficult to set weights in the cost matrix; ii) Resampling has become a practical choice for researchers who are not experts in machine learning domain. Another challenge is generalisation capability is higher in resampling than cost-sensitive approach. Ensemble learning algorithms with any of the above sampling or cost sensitive learning have proven to be the best combination to handle class imbalance (Seiffert et al., 2010; Kuncheva & Rodríguez, 2014).