Employee Attrition Prediction

1. 下圖是我整個分類的過程。（**Fig 1.**）

   其中，在**Preprocessing**的部分，我先把類別的屬性(**BusinessTravel**、**Department**...等)轉為**one-hot**的屬性。以**BusinessTravel**為例，這個屬性有三種值：**Travel_Rarely**、

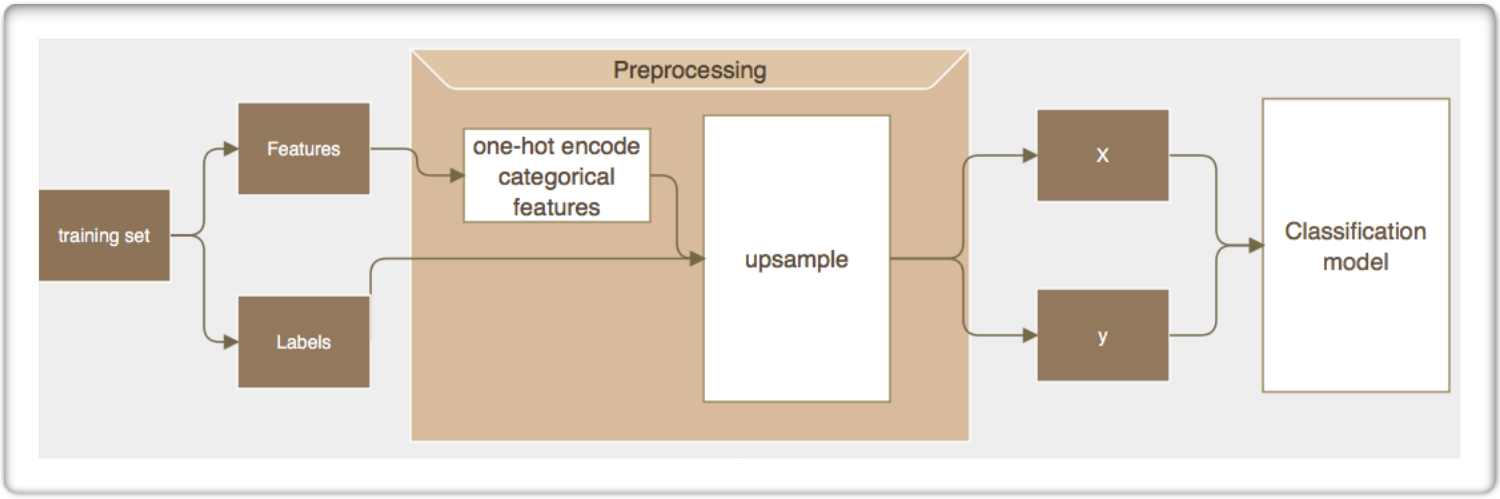

Fig 1. 實驗流程圖

Travel_Frequently以及**Non-Travel**。轉為**one-hot**後會有三種屬性：**BusinessTravel_Travel_Rarely**、**BusinessTravel_Travel_Frequently**以及**BusinessTravel_Non-Travel**，這三種屬性都為**binary**屬性。

把類別屬性都轉為**one-hot**之後會做**upsample**，第二部分會比較不同**upsample**方法的差異，最後選定一個最好的方法，在此部分都是固定用這個方法。

圖中只有寫**training set**的部分，**testing set**也是一樣的流程但不會做**upsample**。

分類模型的部分，我用了三種熱門的分類模型：**Random Forest**、**Support Vector Classifier**以及**XGBoost**。大部分都是用**python**的**sklearn**函式庫實作（除了**XGBoost**是用**xgboost**函式庫）。每個分類模型分別都有實驗有**upsample**以及沒有**upsample**的情況（**Table 1.**）。得到最好的成績是**Random Forest**並且是在有做**upsample**的情況下，在沒做**upsample**的情況下則是**XGBoost**略好。詳細的結果在以下表格。

我有實驗過加入**pca**（**Table 2.**）。不管**component**是多少，效果都更差，在**component = 5**時略好一點但還是比沒做**pca**時更差。所以我就最後決定不做**pca**了。

Model	Upsample	AUROC
Random forest	No	0.59
	Yes	**0.71**
Support vector classifier	No	0.50
	Yes	0.63
XGBoost	No	0.60
	Yes	0.70

Table 1. 各個分類器的效果（包括有無*upsample*的情況）

Model	Upsample	Original	n=2	n=5	n=10
Random forest	No	0.57	0.52	0.57	0.55
	Yes	0.71	0.60	0.61	0.59

Table 2. 以*RF*實驗*PCA*效果（包括有無*upsample*的情況）

另外，在**Random Forest**中我嘗試調了很多參數例如**max_depth**、**bootstrap**、**max_feature**等，也試過或許可以改善**imbalance data**問題的**class_weight**參數，也試過用**feature_importance_**看**feature**的重要程度，刪除最不重要的幾個**feature**再**train**，但預測結果都大同小異甚至在做**upsample**或**pca**後效果更差，所以最後大部分參數都維持**default**。

p.s. random forest每次跑出來的結果都略不同，我盡量挑最好的一次。不過整體的相對**auc**的比較是穩定的。表格內的**auc score**都有在我上傳的**jupyter notebook**內。

2. 關於**upsampling**部分 (這部分都用**Random Forest**（**RF**）做實驗)

我用了三種**upsample**的方法：**Smote**、**Smote+Tomek**(先做**Smote**後刪除**tomek link**)、**Smote+ENN**(先做**Smote**後再做**ENN**)。這部分是用**python**的**imblearn**函式庫實作。下圖是**upsample**前後的兩個類別數量（**Fig 2.**）。

這三種**upsample**方法的比較如下圖（**Fig 3.**）。**Smote + ENN**的效果是最好的，所以我選用這

個方法，包括在第一部分內所有**upsample**的地方都是用此方法。

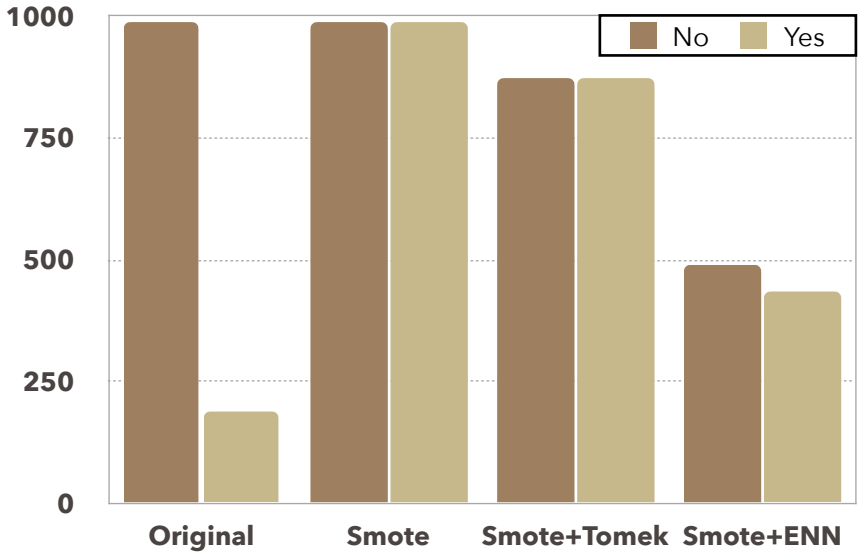

*Fig 2. upsample*前後的數量比較

Model	Original	Smote	Smote + Tomek	Smote + ENN
Random forest	0.59	0.64	0.63	**0.71**

Table 3. 以*RF*實驗*3*種*upsample*效果