데이터 전처리

1 정답 라벨인 'OC', 'sido', 'instkind', 'ownerChange' 는 string으로 분류되어 있는 데이터 형태여서

```
from sklearn.preprocessing import LabelEncoder
```

로 int 로 바꾸었습니다.

2 NaN이 포함되어 있는 행을 삭제하였습니다

```
train=train.dropna(axis=0)
```

3 거의 모든 data가 0으로만 이루어져 있는 'receivableL1', 'receivableL2'를 삭제하였습니다

```
train=train.drop(columns=['receivableL1','receivableL2'])
```

모델학습

	Logistic Regression	KNN	LDA	QDA	Decision Tree
Train으로 예측 accuracy_score (y_train,y_train_pred)	0.9879	0.9759	0.9839	1.0	0.9959
Test로 예측 accuracy_score (y_test,y_test_pred)	0.9285	0.9642	0.8928	0.9642	0.9642

최유경 교수님 기계학습 유튜브 따라보면서 공부했던데 까지 사용했었던 방법들을 전부 사용해보았는데, QDA와 Decision tree가 높게 나왔습니다

특히 RandomForest 를 많이 사용하신 것 같아서 Decision tree에서도 원래는 splitter='best'에서 'random'으로 바꾸어 설정하면 성능이 좋아질까 하여 해보았는데 별로 다른 점이 없었습니다.

Test data set 을 포기한 이유

1 모델학습까지 갔는데, 자꾸 오류가 났습니다
저는 이유가 test의 employee1과 2가 object로, train는 float로 떠서 라고 생각하는데
고치지 못하였습니다

train.info()

48	NCLiabilities2	277 non-null	float64
49	longLoan2	277 non-null	float64
50	netAsset2	277 non-null	float64
51	surplus2	277 non-null	float64
52	employee1	277 non-null	float64
53	employee2	277 non-null	float64
54	ownerChange	277 non-null	int64

dtypes: float64(49), int64(6)

test.info()

48	longLoan2	102 non-null	float64
49	netAsset2	102 non-null	float64
50	surplus2	102 non-null	float64
51	employee1	102 non-null	object
52	employee2	102 non-null	object
53	ownerChange	102 non-null	int64

Test data set 을 포기한 이유

```
1 from sklearn.linear_model import LogisticRegression
2 lg=LogisticRegression(random_state=1)
3 lg.fit(x_train,y_train)
4 y_train_pred=lg.predict(x_train)
5 y_test_pred=lg.predict(test)
```

```
ValueError                                Traceback (most recent call last)
<ipython-input-22-1c96a580cfb4> in <module>()
      3 lg.fit(x_train,y_train)
      4 y_train_pred=lg.predict(x_train)
----> 5 y_test_pred=lg.predict(test)

 5 frames 

/usr/local/lib/python3.7/dist-packages/numpy/core/_asarray.py in asarray(a, dtype, order)
     81
     82     """
---> 83     return array(a, dtype, copy=False, order=order)
     84
     85

ValueError: could not convert string to float: '1,637'
```

2 test에서 NaN이 포함된 행을 지워버렸더니 submission sample과 순서가 맞지 않아서 비어있는 칸들이 생기게 되었습니다