

Lending Club Loan Data Analysis.

DESCRIPTION

Create a model that predicts whether or not a loan will be default using the historical data.

Problem Statement Scenario:

For companies like Lending Club correctly predicting whether or not a loan will be a default is very important. In this project, using the historical data from 2007 to 2015, you have to build a deep learning model to predict the chance of default for future loans. As you will see later this dataset is highly imbalanced and includes a lot of features that makes this problem more challenging.

Domain: Finance

Analysis to be done: Perform data preprocessing and build a deep learning prediction model.

Content:

Dataset columns and definition:

credit.policy:1 if the customer meets the credit underwriting criteria of LendingClub.com, and 0 otherwise.

purpose: The purpose of the loan (takes values "credit_card", "debt consolidation", "educational", "major purchase", "small business", and "all other").

int.rate: The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Borrowers judged by LendingClub.com to be more risky are assigned higher interest rates.

installment: The monthly installments owed by the borrower if the loan is funded.

log.annual.inc: The natural log of the self-reported annual income of the borrower.

dti: The debt-to-income ratio of the borrower (amount of debt divided by annual income).

fico: The FICO credit score of the borrower.

days.with.cr.line: The number of days the borrower has had a credit line.

revol.bal: The borrower's revolving balance (amount unpaid at the end of the credit card billing cycle).

revol.util: The borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available).

inq.last.6mths: The borrower's number of inquiries by creditors in the last 6 months.

delinq.yrs: The number of times the borrower had been 30+ days past due on a payment in the past 2 years.

pub.rec: The borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgments).

Steps to perform:

Perform exploratory data analysis and feature engineering and then apply feature engineering.

Follow up with a deep learning model to predict whether or not the loan will be default using the historical data.

Tasks:

1. Feature Transformation

Transform categorical values into numerical values (discrete)

2. Exploratory data analysis of different factors of the dataset.

3. Additional Feature Engineering

You will check the correlation between features and will drop those features which have a strong correlation
This will help reduce the number of features and will leave you with the most relevant features

4. Modeling

After applying EDA and feature engineering, you are now ready to build the predictive models
In this part, you will create a deep learning model using Keras with TensorFlow backend

Import necessary modules

```
import numpy as np
import pandas as pd

import matplotlib.pyplot as plt
import matplotlib inline
import warnings as sns

warnings.filterwarnings('ignore')
pd.options.display_max_columns = 999
```

Load Data

```
import os
for dirname, _, filenames in os.walk('..\\input\\'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

```
file='../input/loan_data.csv'
```

```
df=pd.read_csv(file)
print("Input Data has ", df.shape[0],"Rows and ",df.shape[1], "Columns", )
```

Input Data has 9578 Rows and 14 Columns

```
df.sample(5)
```

| | credit.policy | purpose | int.rate | installment | log.annual.inc | dti | fico | days.with.cr.line | revol.bal | revol.util | inq.last.6mths | delinq.yrs |
|------|---------------|--------------------|----------|-------------|----------------|-------|------|-------------------|-----------|------------|----------------|------------|
| 4451 | 1 | credit_card | 0.1183 | 106.03 | 10.910186 | 20.96 | 702 | 3630.000000 | 16948 | 83.5 | 0 | 0 |
| 7315 | 1 | debt consolidation | 0.1136 | 789.87 | 11.502875 | 19.79 | 732 | 4650.041667 | 21489 | 52.0 | 0 | 0 |
| 3419 | 1 | all other | 0.1284 | 84.05 | 10.275051 | 17.17 | 692 | 2940.000000 | 1687 | 39.2 | 0 | 0 |
| 4048 | 1 | all other | 0.1284 | 288.95 | 10.308955 | 4.28 | 687 | 4949.958333 | 1602 | 76.3 | 0 | 0 |
| 2921 | 1 | debt consolidation | 0.1474 | 345.37 | 11.736101 | 9.34 | 677 | 5762.000000 | 12403 | 39.8 | 2 | 2 |

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9578 entries, 0 to 9577
Data columns (total 14 columns):
# Column Non-Null Count Dtype
---
0 credit.policy 9578 non-null int64
1 purpose 9578 non-null object
2 int.rate 9578 non-null float64
3 installment 9578 non-null float64
4 log.annual.inc 9578 non-null float64
5 dti 9578 non-null float64
6 fico 9578 non-null int64
7 days.with.cr.line 9578 non-null float64
8 revol.bal 9578 non-null int64
9 revol.util 9578 non-null float64
10 inq.last.6mths 9578 non-null int64
11 delinq.yrs 9578 non-null int64
12 pub.rec 9578 non-null int64
13 not.fully.paid 9578 non-null int64
dtypes: float64(6), int64(7), object(1)
memory usage: 1.7 MB
```

```
df.describe(include='all')
```

| | credit.policy | purpose | int.rate | installment | log.annual.inc | dti | fico | days.with.cr.line | revol.bal |
|--------|---------------|--------------------|-------------|-------------|----------------|-------------|-------------|-------------------|--------------|
| count | 9578.000000 | 7 | 9578.000000 | 9578.000000 | 9578.000000 | 9578.000000 | 9578.000000 | 9578.000000 | 9578.000000 |
| unique | NaN | 7 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| top | NaN | debt_consolidation | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| freq | NaN | 3957 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| mean | 0.804970 | NaN | 0.122640 | 319.089413 | 10.932117 | 12.606679 | 710.846314 | 4560.767197 | 1.691396e+04 |
| std | 0.396245 | NaN | 0.103900 | 163.770000 | 0.614813 | 6.883970 | 37.970537 | 2496.930377 | 3.375619e+04 |
| min | 0.000000 | NaN | 0.060000 | 15.670000 | 7.547502 | 0.000000 | 612.000000 | 178.958333 | 0.000000e+00 |
| 25% | 1.000000 | NaN | 0.103900 | 163.770000 | 10.558414 | 7.212500 | 682.000000 | 2820.000000 | 3.187000e+03 |
| 50% | 1.000000 | NaN | 0.122100 | 268.950000 | 10.928884 | 12.665000 | 707.000000 | 4139.958333 | 8.596000e+03 |
| 75% | 1.000000 | NaN | 0.140700 | 432.762500 | 11.291293 | 17.950000 | 737.000000 | 5730.000000 | 1.824950e+04 |
| max | 1.000000 | NaN | 0.216400 | 940.140000 | 14.528354 | 29.900000 | 827.000000 | 17639.958330 | 1.207359e+06 |

```
df.describe().transpose()
```

| | count | mean | std | min | 25% | 50% | 75% | max |
|-------------------|--------|--------------|--------------|------------|-------------|-------------|--------------|--------------|
| credit.policy | 9578.0 | 0.804970 | 0.396245 | 0.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000e+00 |
| int.rate | 9578.0 | 0.122640 | 0.026847 | 0.060000 | 0.103900 | 0.122100 | 0.140700 | 2.164000e+01 |
| installment | 9578.0 | 319.089413 | 207.071301 | 15.670000 | 163.770000 | 268.950000 | 432.762500 | 9.401400e+02 |
| log.annual.inc | 9578.0 | 10.932117 | 0.614813 | 7.547502 | 10.558414 | 10.928884 | 11.291293 | 1.528354e+01 |
| dti | 9578.0 | 12.606679 | 6.883970 | 0.000000 | 7.212500 | 12.665000 | 17.950000 | 2.960000e+01 |
| fico | 9578.0 | 710.846314 | 37.970537 | 612.000000 | 682.000000 | 707.000000 | 737.000000 | 8.270000e+02 |
| days.with.cr.line | 9578.0 | 4560.767197 | 2496.930377 | 178.958333 | 2820.000000 | 4139.958333 | 5730.000000 | 1.763996e+04 |
| revol.bal | 9578.0 | 16913.963876 | 33756.189557 | 0.000000 | 3187.000000 | 8596.000000 | 18249.500000 | 1.207359e+06 |
| revol.util | 9578.0 | 46.799236 | 29.014417 | 0.000000 | 22.600000 | 46.300000 | 70.900000 | 1.190000e+02 |
| inq.last.6mths | 9578.0 | 1.577469 | 2.200245 | 0.000000 | 0.000000 | 1.000000 | 2.000000 | 3.300000e+01 |
| delinq.yrs | 9578.0 | 0.163708 | 0.546215 | 0.000000 | 0.000000 | 0.000000 | 1.000000 | 3.000000e+01 |
| pub.rec | 9578.0 | 0.062122 | 0.262126 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 5.000000e+00 |
| not.fully.paid | 9578.0 | 0.160054 | 0.366676 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 1.000000e+00 |

Check for missing values or null values

```
df.isnull().sum().sum()
```

0

```
NaN = pd.concat((df.isnull().sum(),axis=1),keys=['Count of NaNs'])
NaNs.sample(14)
```

```
Count of NaNs
```

| | Count of NaNs |
|-------------------|---------------|
| revol.util | 0 |
| fico | 0 |
| installment | 0 |
| delinq.yrs | 0 |
| not.fully.paid | 0 |
| pub.rec | 0 |
| inq.last.6mths | 0 |
| days.with.cr.line | 0 |
| log.annual.inc | 0 |
| revol.bal | 0 |
| credit.policy | 0 |
| int.rate | 0 |
| purpose | 0 |
| dti | 0 |

```
NaN[NaNs.sum(axis=1)>0]
```

```
Count of NaNs
```

It is good that there is no missing values or no null values

Now let us have a look at the data type of all the variables present in the input dataset

```
# Now let us have a look at the data type of all the variables present in the training dataset
dtype_data=df.dtypes.reset_index()
dtype_data.columns = ['Column','Column Type']
dtype_data.groupby('Column').aggregate('count').reset_index()
```

| Column Type | Count |
|-------------|-------|
| int64 | 7 |
| float64 | 6 |
| object | 1 |

So 7 columns are integers with 1 object column and 6 float column

Analyze the column types of input data

```
# possible data types in pandas
numeric=['int16','int32','int64','float16','float32','float64'] # numeric
objects = ['O']

df_num = df.select_dtypes(include=numeric)
df_cat = df.select_dtypes(include=objects)

print(df_cat.columns)
print(df_num.columns)
```

```
Index(['purpose'], dtype='object')
Index(['credit.policy', 'int.rate', 'installment', 'log.annual.inc', 'dti',
       'fico', 'days.with.cr.line', 'revol.bal', 'revol.util',
       'inq.last.6mths', 'delinq.yrs', 'pub.rec', 'not.fully.paid',
       dtype='object')
```

Check the label "not.fully.paid" distribution in the dataset.

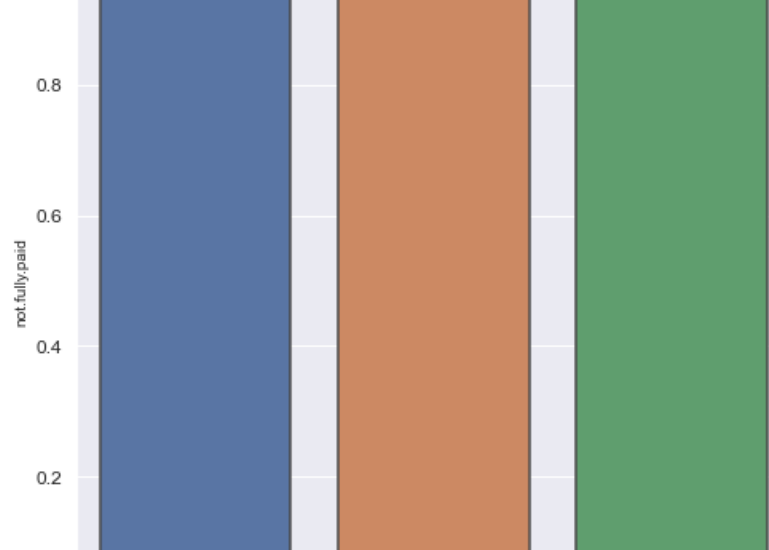
```
df['not.fully.paid'].isnull().mean()
df.groupby('not.fully.paid')['not.fully.paid'].count()/len(df)
```

```
not.fully.paid
0    0.839946
1    0.160054
Name: not.fully.paid, dtype: float64
```

```
sns.set_style('darkgrid')
```

```
sns.countplot(x='not.fully.paid', data=df)
```

```
<AxesSubplot:xlabel='not.fully.paid', ylabel='count'>
```



The above shows, This dataset is highly imbalanced and includes features that make this problem more challenging. If we do model training with this data, the prediction will be biased since the "not.fully.paid = 0" has 83.9% filled, and only 16% is the "not.fully.paid=1"

To handle the imbalance of target variable "not.fully.paid" Oversampling is used when the quantity of data is insufficient. It tries to balance the dataset by increasing the size of rare samples.

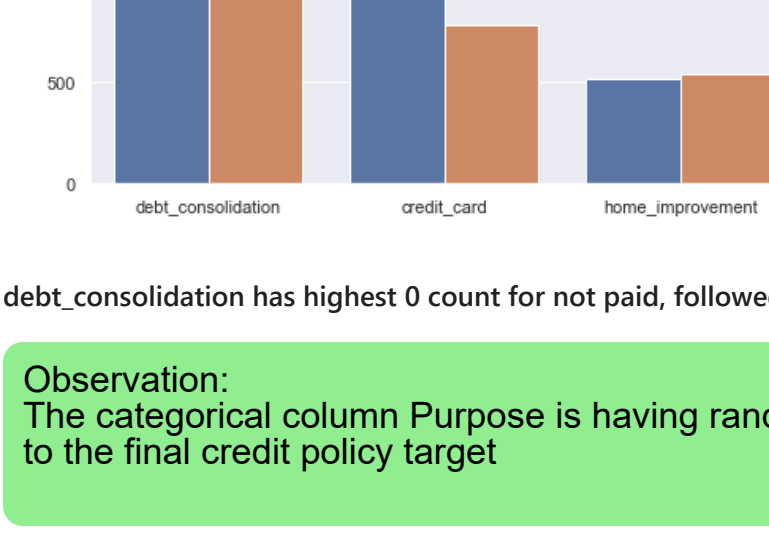
Over-sample the minority class with or w/o replacement by making the number of positive and negative examples equal. We'll add around 6500 samples from the training data set with this strategy. It's a lot more computationally expensive than under-sampling. Also, it's more prone to overfitting due to repeated examples.

```
count_class0,count_class1 = df['not.fully.paid'].value_counts()
df_0 = df[df['not.fully.paid'] == 0]
df_1 = df[df['not.fully.paid'] == 1]
df_1_over = df_1.sample(count_class0, replace=True)
df_over = pd.concat((df_0, df_1_over), axis=0)

print(df_over['not.fully.paid'].value_counts())
```

```
sns.set_style('darkgrid')
sns.countplot(x='not.fully.paid', data=df_over)
```

```
Random over-sampling:
0    8045
1    8045
Name: not.fully.paid, dtype: int64
<AxesSubplot:xlabel='not.fully.paid', ylabel='count'>
```



```
df_over = df_over.reset_index()
df_over[df_over.index.duplicated()]
```

```
index credit.policy purpose int.rate installment log.annual.inc dti fico days.with.cr.line revol.bal revol.util inq.last.6mths delinq.yrs
```

Categorical Variables:

```
df[0:df.shape[0]-1,:]
```

```
Count Column Type
1 purpose object
```

So purpose is the only object column

```
df_over[1:'purpose'].sample(10)
```

```
purpose
```

```
6573 all other
```

```
6215 credit card
```

```
12066 all other
```

```
6034 debt consolidation
```

```
10562 major purchase
```

```
1561 small business
```

```
283 credit card
```

```
13139 all other
```

```
8030 debt consolidation
```

```
2897 all other
```

The column purpose is categorical

Looking into categorical feature

```
# display distinct value for each categorical feature
df_cat = df_over.select_dtypes(include=objects)
for col_name in df_cat.columns:
    print('The unique values in ' + col_name + ' are: ', df_cat[col_name].unique())
    print(df_cat[col_name].unique())
```

```
The unique values in purpose are: 7
['debt_consolidation', 'credit_card', 'home_improvement', 'small_business',
 'major_purchase', 'all other', 'educational']
```

```
df_over['purpose'].value_counts()
```

```
debt_consolidation 6460
```

```
all other 4037
```

```
credit_card 1302
```

```
small_business 1352
```

```
home_improvement 1062
```

```
major_purchase 643
```

```
educational 634
```

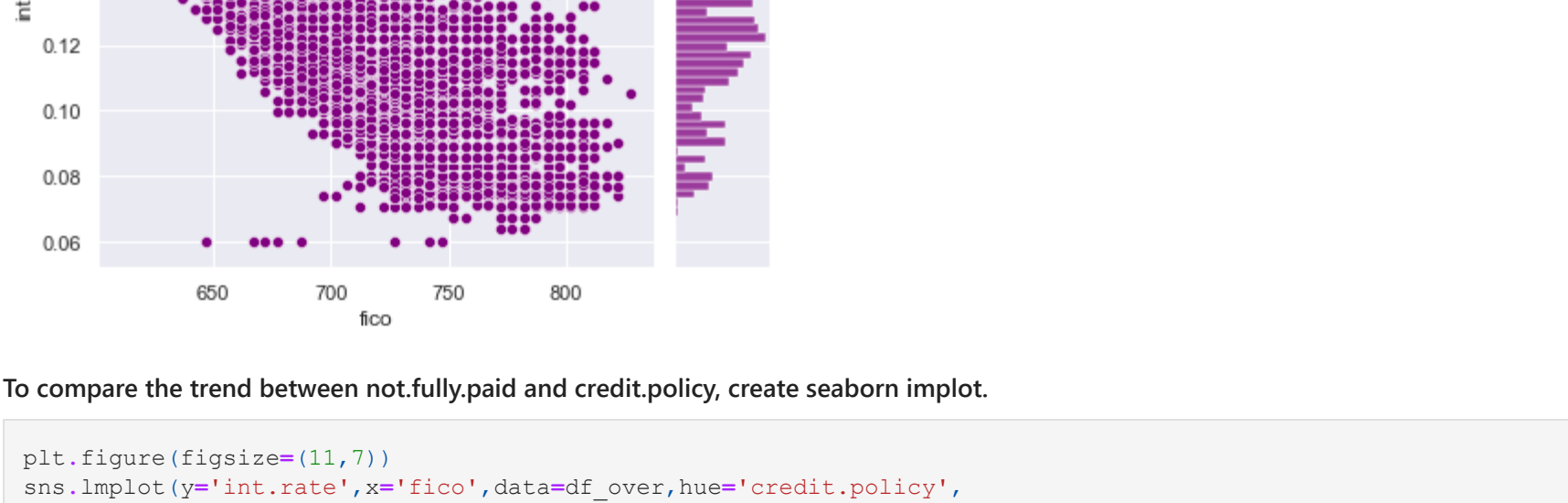
```
Name: purpose, dtype: int64
```

Analyze the not.fully.paid - for cat column purpose

* To assess the usefulness of categorical feature we shall use boxplot

```
# value of credit.policy(target variable) change for categorical column purpose ....
cols=['purpose']

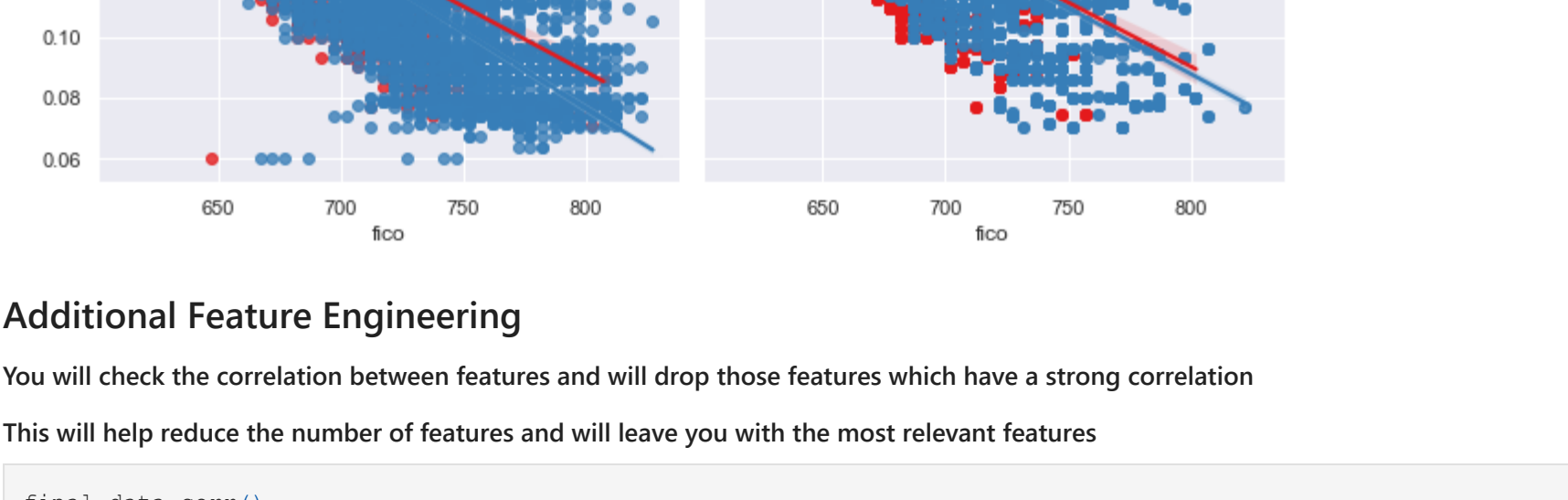
for col in cols:
    plt.figure(figsize=(18,7,8,27))
    sns.boxplot(x=col,y='not.fully.paid',data=df_over)
    plt.xlabel(col,fontsize=10)
    plt.ylabel('figure.figsize':(18,7,8,27))
    plt.xticks(fontsize=12)
    plt.yticks(fontsize=12)
```



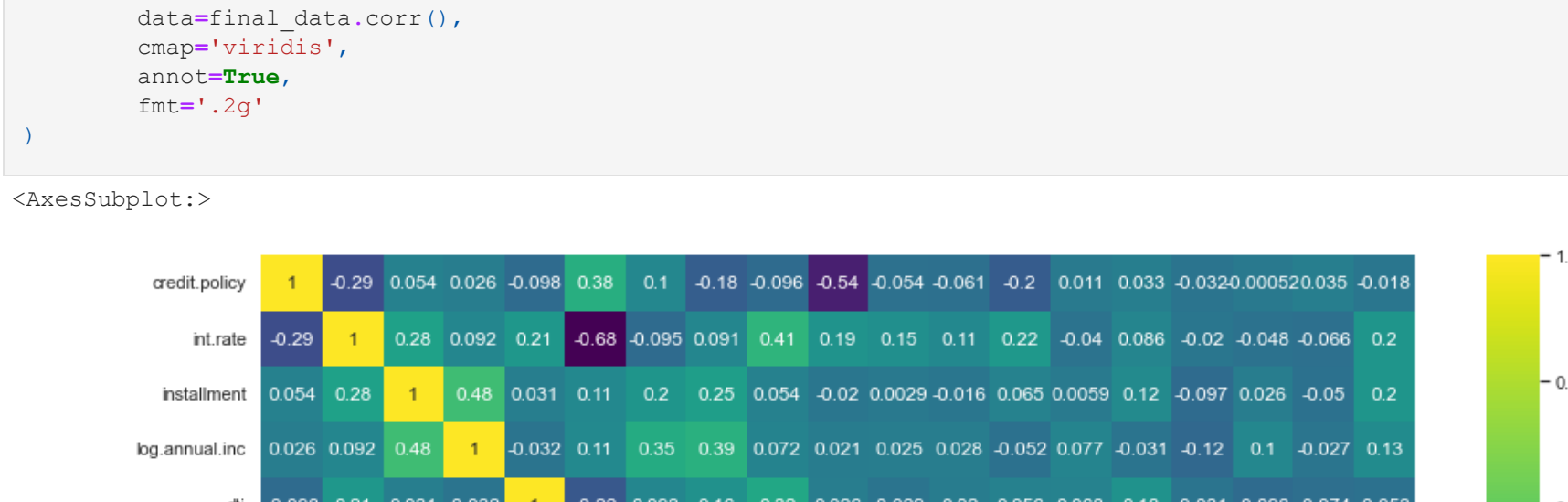
Boxplot shows not.fully.paid data distribution is 1 or 0 and spread across all the categories of purpose.

Let us see the count of different purpose

```
sns.set_theme(style='darkgrid')
sns.set(rc={'figure.figsize':(18,7,8,27)})
ax = sns.countplot(x='purpose', data=df_over)
```



Countplot showing debt consolidation purpose have highest no. of purpose for not paid with lowest being education



debt consolidation has highest Count for not paid, followed by all other and credit card purpose

Observation:

The categorical column Purpose is having random distribution for every category and would contribute to the final credit policy target

Feature Transformation

Transform categorical values into numerical values (discrete)

```
col_fea = ['purpose']
final_data = pd.get_dummies(df_over,columns=col_fea,drop_first=True)
final_data = final_data.drop(['index'],axis=1)
final_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16090 entries, 0 to 16089
Data columns (total 19 columns):
# Column Non-Null Count Dtype
---
0 credit.policy 16090 non-null int64
1 int.rate 16090 non-null float64
2 installment 16090 non-null float64
3 log.annual.inc 16090 non-null float64
4 dti 16090 non-null float64
5 fico 16090 non-null int64
6 days.with.cr.line 16090 non-null float64
7 revol.bal 16090 non-null int64
8 revol.util 16090 non-null float64
9 inq.last.6mths 16090 non-null int64
10 delinq.yrs 16090 non-null int64
11 pub.rec 16090 non-null int64
12 not.fully.paid 16090 non-null int64
13 purpose_credit_card 16090 non-null uint8
14 purpose_debt_consolidation 16090 non-null uint8
15 purpose_educational 16090 non-null uint8
16 purpose_home_improvement 16090 non-null uint8
17 purpose_major_purchase 16090 non-null uint8
18 purpose_small_business 16090 non-null uint8
dtypes: float64(6), int64(7), uint8(6)
memory usage: 1.7 MB
```

Exploratory data analysis of different factors of the dataset.

Create a histogram of two FICO distributions on top of each other, one for each credit.policy outcome.

Let's see a similar chart for "not.fully.paid" column.

Now, check the dataset group by loan purpose. Create a countplot with the color hue defined by not.fully.paid.

The next visual we will pull part of EDA in this dataset is the trend between FICO score and interest rate.

To compare the trend between not.fully.paid and credit policy, create seaborn implot.

| | | Column | Non-Null Count | Dtype |
|--|----------------------------|--------|----------------|---------|
| Data columns (total 35 columns): | | | | |
| 0 | credit_policy | | 16090 non-null | int64 |
| 1 | int_rate | | 16090 non-null | float64 |
| 2 | annual_inc | | 16090 non-null | float64 |
| 3 | dti | | 16090 non-null | float64 |
| 4 | fico | | 16090 non-null | int64 |
| 5 | inq_last_6mths | | 16090 non-null | int64 |
| 6 | delinq_2yrs | | 16090 non-null | int64 |
| 7 | pub_rec | | 16090 non-null | int64 |
| 8 | totally_paid | | 16090 non-null | int64 |
| 9 | purpose_credit_card | | 16090 non-null | int64 |
| 10 | purpose_debt_consolidation | | 16090 non-null | int64 |
| 11 | purpose_educational | | 16090 non-null | int64 |
| 12 | purpose_home_improvement | | 16090 non-null | int64 |
| 13 | purpose_major_purchase | | 16090 non-null | int64 |
| 14 | purpose_small_business | | 16090 non-null | int64 |
| dtypes: float64(3), int64(6), uint8(6) | | | | |
| memory usage: 1.2 MB | | | | |

Important Variables:

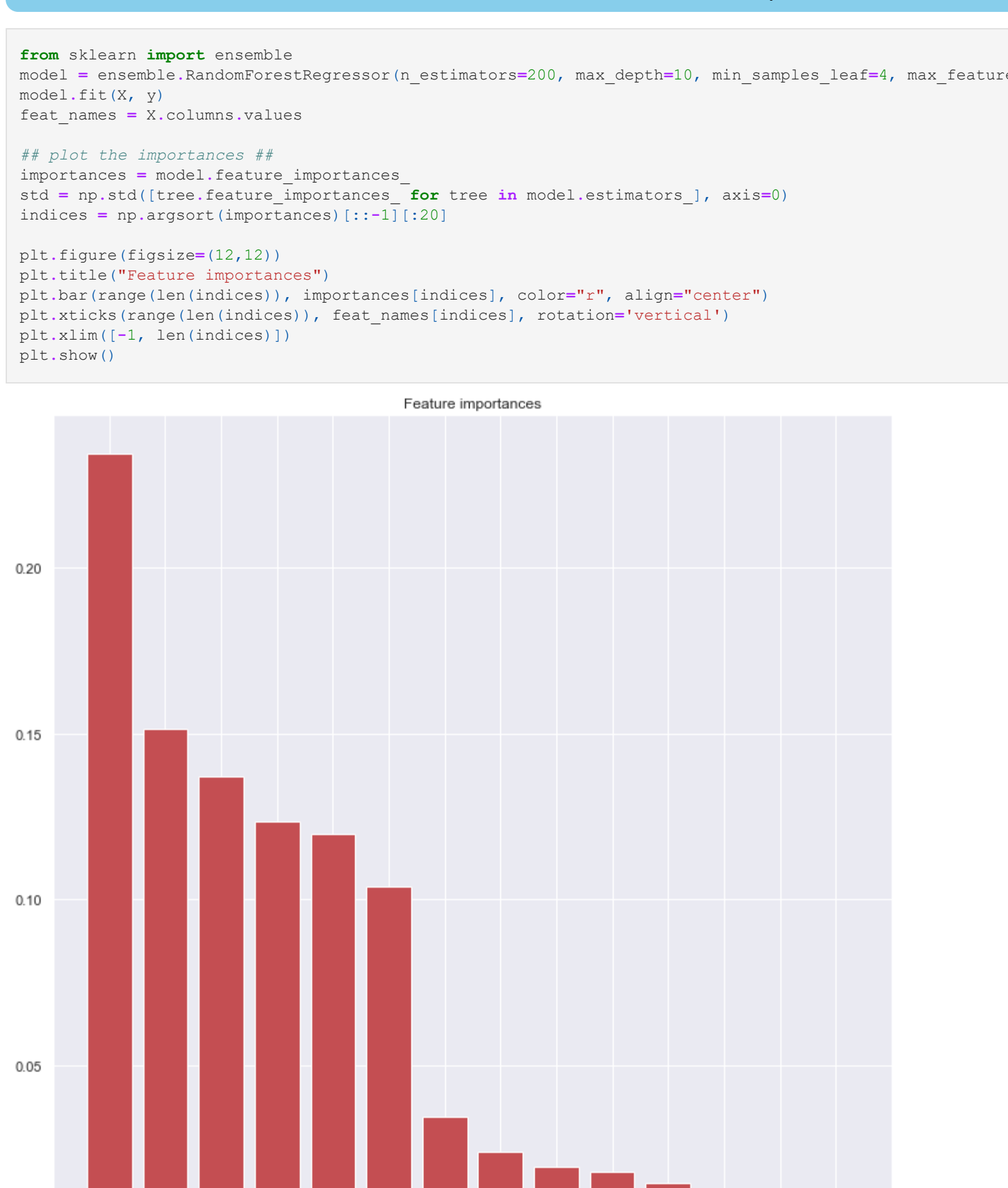
Now let us run and xgboost model to get the important variables.

```
In [208]:
from sklearn.preprocessing import LabelEncoder
from sklearn import preprocessing
import xgboost as xgb
import numpy as np
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=101)
le = LabelEncoder()
X_train = le.fit_transform(X_train)
X_test = le.transform(X_test)
def xgb_r2_score(preds, dtrain):
    labels = dtrain.get_label()
    return 'r2', r2_score(labels, preds)
xgb_params = {
    'eta': 0.05,
    'max_depth': 6,
    'subsample': 0.7,
    'colsample_bytree': 0.7,
    'objective': 'reg:linear',
    'silent': 1
}
dtrain = xgb.DMatrix(X, y, feature_names=X.columns.values)
cv_results = xgb.cv(dtrain, xgb_params, num_boost_round=100, feval=xgb_r2_score, maximize=True)
fig = plt.figure(figsize=(12,12))
xgb.plot_importance(model, max_num_features=50, height=0.8, ax=fig)
plt.show()
```

(2043,126) WARNING: C:\Users\Administrator\workspace\xgboost-win64_release_1.4.0\src\objective\regression_obj.o: util::reglinear is now deprecated in favor of reglinearquadrant.

(2043,126) WARNING: C:\Users\Administrator\workspace\xgboost-win64_release_1.4.0\src\learner.cc:573: Parameters: { "silent" } might not be used.

dti may not be accurate due to some parameters are only used in language bindings but passed down to XGBoost core. Or some parameters are not used but slip through this verification. Please open an issue if you find above cases.



Let us also build a Random Forest model and check the important variables.

```
In [209]:
from sklearn import ensemble
from sklearn.ensemble import RandomForestRegressor
n_estimators=200, max_depth=10, min_samples_leaf=4, max_features=0.2, n_jobs=-1
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=101)
model = RandomForestRegressor(n_estimators=200, max_depth=10, min_samples_leaf=4, max_features=0.2, n_jobs=-1)
model.fit(X_train, y_train)
importances = model.feature_importances_
indices = np.argsort(importances)[::-1][1:20]
plt.figure(figsize=(12,12))
plt.title("Feature Importances")
plt.bar(range(len(indices)), importances[indices], color="r", align="center")
plt.xticks(range(len(indices)), feature_names[indices], rotation=45)
plt.xlim(-1, len(indices))
plt.show()
```



Deep Learning Model Implementation

Do the train test split and fit the model with the data shape we created above, since there are now 14 features, the first layer of the neural network is created with 14 nodes.

```
In [210]:
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout
from tensorflow.keras.callbacks import EarlyStopping
from tensorflow.keras.models import load_model
from sklearn.metrics import confusion_matrix, classification_report
```

In [210]:

```
to_train = final_data_fe[final_data_fe['not_fully_paid'] != 1].index[0:11]
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state = 101)
```

```
scaler = MinMaxScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

```
model = Sequential()
```

```
model.add(Dense(14, activation='relu'))
```

```
model.add(Dense(10, activation='relu'))
```

```
model.add(Dense(5, activation='relu'))
```

```
model.add(Dense(1, activation='sigmoid'))
```

```
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])
```

```
early_stop = EarlyStopping(monitor='val_loss', mode='min', verbose=1, patience=25)
```

```
model.fit(X_train, y_train, batch_size=256, validation_data=(X_test, y_test), callbacks=[early_stop])
```

```
Epoch 1/200: loss=0.6902, accuracy=0.5420, val_loss=0.6897, val_accuracy=0.5823
```

```
Epoch 2/200: loss=0.6804, accuracy=0.5987, val_loss=0.6747, val_accuracy=0.6004
```

```
Epoch 3/200: loss=0.6684, accuracy=0.6065, val_loss=0.6651, val_accuracy=0.6012
```

```
Epoch 4/200: loss=0.6593, accuracy=0.6109, val_loss=0.6591, val_accuracy=0.6035
```

```
Epoch 5/200: loss=0.6547, accuracy=0.6150, val_loss=0.6565, val_accuracy=0.6058
```

```
Epoch 6/200: loss=0.6515, accuracy=0.6181, val_loss=0.6555, val_accuracy=0.6111
```

```
Epoch 7/200: loss=0.6497, accuracy=0.6202, val_loss=0.6549, val_accuracy=0.6099
```

```
Epoch 8/200: loss=0.6477, accuracy=0.6188, val_loss=0.6532, val_accuracy=0.6062
```

```
Epoch 9/200: loss=0.6461, accuracy=0.6212, val_loss=0.6526, val_accuracy=0.6124
```

```
Epoch 10/200: loss=0.6450, accuracy=0.6240, val_loss=0.6519, val_accuracy=0.6114
```

```
Epoch 11/200: loss=0.6439, accuracy=0.6242, val_loss=0.6524, val_accuracy=0.6178
```

```
Epoch 12/200: loss=0.6436, accuracy=0.6229, val_loss=0.6508, val_accuracy=0.6085
```

```
Epoch 13/200: loss=0.6426, accuracy=0.6253, val_loss=0.6510, val_accuracy=0.6167
```

```
Epoch 14/200: loss=0.6422, accuracy=0.6258, val_loss=0.6507, val_accuracy=0.6157
```

```
Epoch 15/200: loss=0.6419, accuracy=0.6281, val_loss=0.6500, val_accuracy=0.6145
```

```
Epoch 16/200: loss=0.6415, accuracy=0.6247, val_loss=0.6502, val_accuracy=0.6126
```

```
Epoch 17/200: loss=0.6411, accuracy=0.6257, val_loss=0.6499, val_accuracy=0.6138
```

```
Epoch 18/200: loss=0.6400, accuracy=0.6302, val_loss=0.6494, val_accuracy=0.6132
```

```
Epoch 19/200: loss=0.6398, accuracy=0.6298, val_loss=0.6495, val_accuracy=0.6114
```

```
Epoch 20/200: loss=0.6394, accuracy=0.6282, val_loss=0.6496, val_accuracy=0.6114
```

```
Epoch 21/200: loss=0.6392, accuracy=0.6291, val_loss=0.6495, val_accuracy=0.6107
```

```
Epoch 22/200: loss=0.6388, accuracy=0.6294, val_loss=0.6491, val_accuracy=0.6107
```

```
Epoch 23/200: loss=0.6386, accuracy=0.6297, val_loss=0.6499, val_accuracy=0.6165
```

```
Epoch 24/200: loss=0.6387, accuracy=0.6263, val_loss=0.6494, val_accuracy=0.6149
```

```
Epoch 25/200: loss=0.6385, accuracy=0.6302, val_loss=0.6487, val_accuracy=0.6116
```

```
Epoch 26/200: loss=0.6386, accuracy=0.6271, val_loss=0.6485, val_accuracy=0.6134
```

```
Epoch 27/200: loss=0.6378, accuracy=0.6308, val_loss=0.6484, val_accuracy=0.6134
```

```
Epoch 28/200: loss=0.6378, accuracy=0.6304, val_loss=0.6501, val_accuracy=0.6159
```

```
Epoch 29/200: loss=0.6384, accuracy=0.6256, val_loss=0.6484, val_accuracy=0.6116
```

```
Epoch 30/200: loss=0.6370, accuracy=0.6294, val_loss=0.6494, val_accuracy=0.6107
```

```
Epoch 31/200: loss=0.6372, accuracy=0.6297, val_loss=0.6478, val_accuracy=0.6130
```

```
Epoch 32/200: loss=0.6368, accuracy=0.6280, val_loss=0.6480, val_accuracy=0.6147
```

```
Epoch 33/200: loss=0.6369, accuracy=0.6312, val_loss=0.6486, val_accuracy=0.6201
```

```
Epoch 34/200: loss=0.6370, accuracy=0.6306, val_loss=0.6477, val_accuracy=0.6153
```

```
Epoch 35/200: loss=0.6375, accuracy=0.6276, val_loss=0.6479, val_accuracy=0.6147
```

```
Epoch 36/200: loss=0.6369, accuracy=0.6311, val_loss=0.6474, val_accuracy=0.6136
```

```
Epoch 37/200: loss=0.6359, accuracy=0.6330, val_loss=0.6488, val_accuracy=0.6174
```

```
Epoch 38/200: loss=0.6358, accuracy=0.6303, val_loss=0.6475, val_accuracy=0.6172
```

```
Epoch 39/200: loss=0.6357, accuracy=0.6314, val_loss=0.6477, val_accuracy=0.6188
```

```
Epoch 40/200: loss=0.6355, accuracy=0.6302, val_loss=0.6473, val_accuracy=0.6163
```

```
Epoch 41/200: loss=0.6352, accuracy=0.6311, val_loss=0.6471, val_accuracy=0.6209
```

```
Epoch 42/200: loss=0.6347, accuracy=0.6295, val_loss=0.6473, val_accuracy=0.6203
```

```
Epoch 43/200: loss=0.6346, accuracy=0.6311, val_loss=0.6468, val_accuracy=0.6163
```

```
Epoch 44/200: loss=0.6346, accuracy=0.6289, val_loss=0.6469, val_accuracy=0.6184
```

```
Epoch 45/200: loss=0.6350, accuracy=0.6332, val_loss=0.6469, val_accuracy=0.6188
```

```
Epoch 46/200: loss=0.6344, accuracy=0.6332, val_loss=0.6471, val_accuracy=0.6232
```

```
Epoch 47/200: loss=0.6350, accuracy=0.6309, val_loss=0.6471, val_accuracy=0.6248
```

```
Epoch 48/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 49/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 50/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 51/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 52/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 53/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 54/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 55/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 56/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 57/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 58/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 59/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 60/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 61/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 62/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 63/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 64/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 65/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 66/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 67/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 68/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 69/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 70/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 71/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 72/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 73/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 74/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 75/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 76/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 77/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 78/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 79/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 80/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 81/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 82/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 83/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 84/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 85/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 86/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 87/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 88/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 89/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 90/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 91/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 92/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 93/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 94/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 95/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

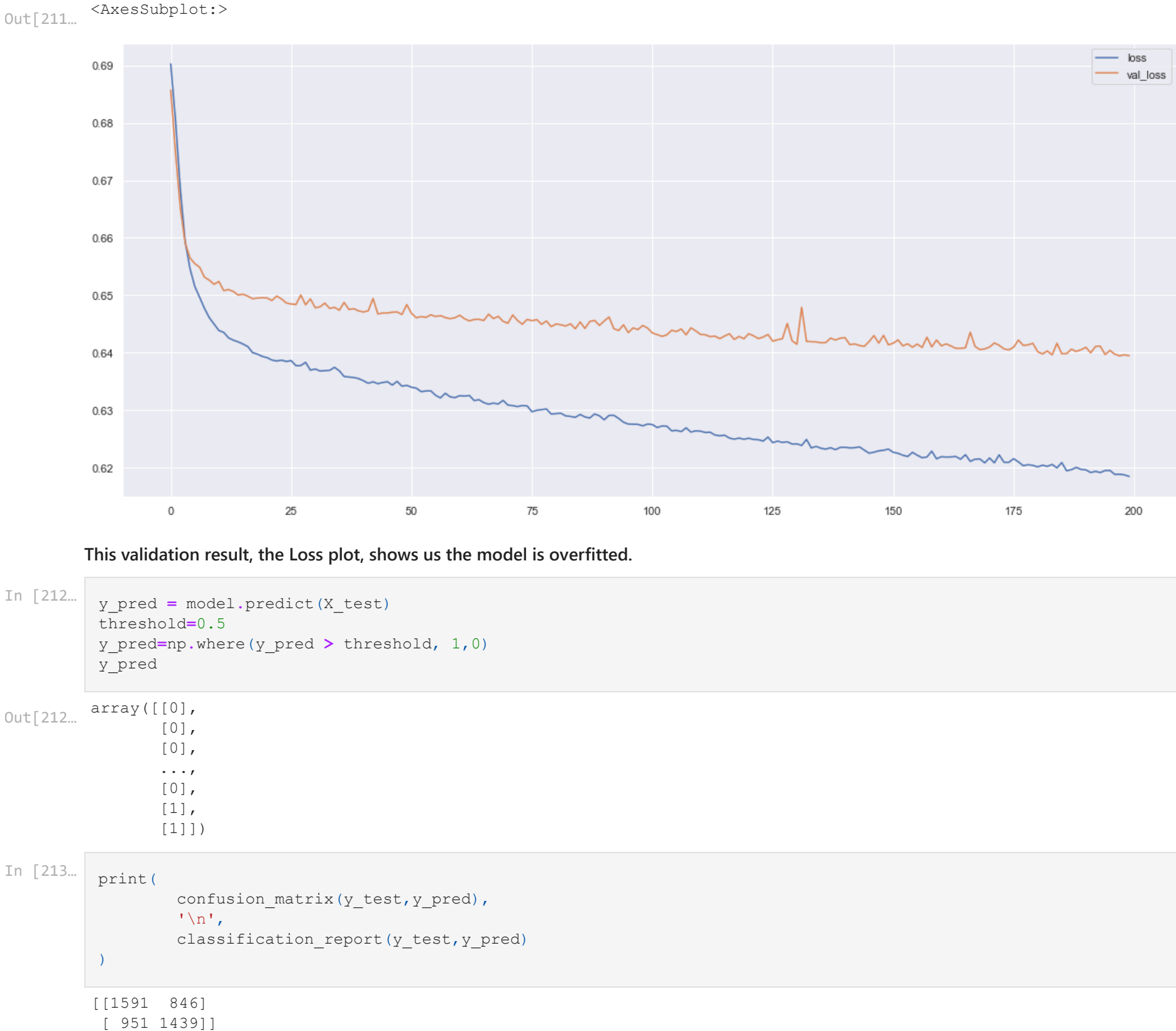
```
Epoch 96/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 97/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 98/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 99/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

```
Epoch 100/200: loss=0.6342, accuracy=0.6332, val_loss=0.6467, val_accuracy=0.6180
```

| | precision | recall | f1-score | support |
|---|-----------|--------|----------|---------|
| 0 | 0.63 | 0.65 | 0.64 | 2437 |
| 1 | 0.63 | 0.60 | 0.62 | 2300 |

| | 1 | 0.63 | 0.60 | 0 |
|--------------|------|------|------|---|
| accuracy | | | | 0 |
| macro avg | 0.63 | 0.63 | | 0 |
| weighted avg | 0.63 | 0.63 | | 0 |

The model's overall f1-score for accuracy is

Mod
Two w
Add D

```
model_new.add(Dropout(0.2))

model_new.add(
    Dense(10, activation='relu')
)

model_new.add(Dropout(0.2))

model_new.add(
    Dense(5, activation='relu')
)
```

```
model_new.add(Dropout(0.2))
```

```
model_new.add(
    Dense(1, acti
)
```

```
model_new.compile(
    optimizer='adam',
    loss='binary_crossentropy',
    metrics=['binary_accuracy'])

model_new.fit(
    X_train,
    y_train,
    epochs=200,
    batch_size=32,
    validation_data=(X_test, y_test),
    callbacks=[early_stop])
```

```
44/44 [=====]
9 - val_binary_accuracy: 0.9999
Epoch 2/200
```

```
44/44 [=====] -
7 - val_binary_accuracy: 0.6029
```

```
Epoch 3/2000
44/44 [=====] - 0s 3ms/step - loss: 0.6846 - binary_accuracy: 0.5577 - val_loss: 0
2 - val_binary_accuracy: 0.5393
Epoch 4/2000
44/44 [=====] - 0s 3ms/step - loss: 0.6803 - binary_accuracy: 0.5605 - val_loss: 0
4 - val_binary_accuracy: 0.6024
Epoch 5/2000
44/44 [=====] - 0s 3ms/step - loss: 0.6749 - binary_accuracy: 0.5760 - val_loss: 0
8 - val_binary_accuracy: 0.6068
Epoch 6/2000
```

[illegible]

By changing the cut-off line to 0.2 (default is 0.5), we have dramatically brought down the Type 2 error.

Conclusion

When building the Neural Network, the most difficult part is the Sequential Model because there are many different options available in building the layers. The way how to come up with the optimized number of layers and nodes are remaining challenging.