Homework 4

Course: Introduction to applied data science (PHYS247) Spring 2020
TA: Nima Chartab Due Date: June 1, 11:59 p.m.

Problem 1: LendingClub

LendingClub is an American peer-to-peer lending company, headquartered in San Francisco, California. We want to build a predictive model that can predict whether or not a borrower will pay back their loan. To create such model, you are given a data set ("LendingClub.csv") with detailed information of the past borrowers. The dataset has 23 columns which are described as below:

loan_amnt: The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
term: The number of payments on the loan. Values are in months and can be either 36 or 60.
int_rate: Interest Rate on the loan
installment: The monthly payment owed by the borrower if the loan originates.
grade: LC assigned loan grade
sub_grade: LC assigned loan subgrade
home_ownership: The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
annual_inc: The self-reported annual income provided by the borrower during registration.
verification_status: Indicates if income was verified by LC, not verified, or if the income source was verified
loan_status: Current status of the loan
purpose: A category provided by the borrower for the loan request.
dti: A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
earliest_cr_line: The month the borrower's earliest reported credit line was opened
open_acc: The number of open credit lines in the borrower's credit file.
pub_rec: Number of derogatory public records
revol_bal: Total credit revolving balance
revol_util: Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
total_acc: The total number of credit lines currently in the borrower's credit file
initial_list_status: The initial listing status of the loan. Possible values are W and F
application_type: Indicates whether the loan is an individual application or a joint application with two co-borrowers
mort_acc: Number of mortgage accounts.
pub_rec_bankruptcies: Number of public record bankruptcies
addr_state: The state provided by the borrower in the loan application

a) First perform data cleaning. Convert all categorical variables to numeric variables. Extract the zip-code information from the borrowers' address and consider that as one of your features.
b) Split the data to training (70%) and test(30%) sample.
c) Use sklearn.preprocessing.MinMaxScaler function to normalize all the features for both training and test sample.
d) Build a sequential model using neural-network library Keras. Use four Dense layers (including input and output layers) along with dropout regularization to reduce overfitting. You need to decide on the number of neurons in each layer.
d) Evaluate the performance of your model.