EE 660

Project Proposal: Type 1

Posted: Sat., 10/29/2022 Due: Mon., 11/7/2022, 11:59 PM PST

**This proposal form is for Type 1 projects: Solve a ML problem by implementing a ML system of your own design, using real-world data.**

**Please fill in both the Project Proposal form (pp. 1-2) and the Dataset Information Form (p. 3)**. **This is required of everyone (each team submits one Project Proposal with all their names on it, including one dataset form for each dataset they will use).** \*All fields except “other comments” are required. In each field, replace instructions (black text) with your descriptions. Preferred format is to enter your answers into the Word version of this form, then convert to pdf before submission. If you prefer to use another app instead of Word, then submit a typed version with each field labeled with its title (“Dataset”, etc.), and submit as a pdf file.

The primary purpose of this proposal is to give you some feedback on your project topic and plans. This proposal will be graded and will count ~10% toward your total project grade. The grading of this proposal will be primarily based on whether you put in a reasonable effort and whether the content makes good technical sense.

|  |
| --- |
| The Bank Client Subscribed to Term Deposit Prediction |
| \*Project team: Your name(s) and email address(es) |
| Lijing Yang: lijingy@usc.edu |
| \*Clear statement of the problem and/or goals. |
| I will use different classification methods such as KNN, Naive Bayes, SVM and CART in this regression problem to compare the different performance of different classifiers on this dataset. In addition to that I would also compare which features are the most predictive. |
| \*Clear statement of how your problem could be related with Transfer Learning (TL), Semi-supervised Learning (SSL) or other extension tasks. |
| Labels are precious, so we often cannot rely too much on labeled data. The data I chose in this project is fully labeled data, so I will select a part of the data to remove its labels as unlabeled data and use SSL to train the model. I will use self-training or other training methods on unlabeled data and check the model's accuracy by adjusting the proportion of unlabeled data. And compare the model using the SSL method with the SL model. |
| \*A plan of preprocessing and feature extraction (if applicable) |
| Since it is a binary classification, I will check if the data is balanced. If the data is unbalanced, my first step in preprocessing will be to balance the data first. I have 20 features in my dataset, such as age, job, marital, education, etc. I will use these 20 features to fit and then test and remove the completely useless features or contribute too little to the classification. |
| \*A plan of your approach |
| I will first test several classifiers learned in 660, such as KNN, CART, etc., and then use some classifiers learned in 559, such as SVM and Naive Bayes. Because there are currently 20 features, I think this feature amount is just right, so I will not increase or decrease my data's features. Since this is a classification problem, I will analyze the performance of a model by looking for training and testing confusion matrices. |
| \*A description of any other work of yours that is related to your proposed class project |
| “None” |
| \*If yours is a team project, roughly describe how work will be divided |
| I’m doing this project by myself |
| Other Comments |
| There are no comments now, but I believe I will have Comments in the process of doing it. |

EE 660 Dataset Information Form Fall 2022

*Include one form for each dataset you plan to use. (For each dataset’s form, you may continue onto an additional page if necessary.)*

\*Dataset or competition title: Bank Marketing Data Set

\*Link: <https://archive.ics.uci.edu/ml/datasets/Bank+Marketing>

**\*Problem type**: classification/logistic regression

**\*Brief description of dataset and problem domain**: The data is related to direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact with the same client was required to assess if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

**\*Number of data points**: 41188

**\*Number of features or input variables**: 20

\*Feature or input-variable types: numeric (10 variables, 5 int, 5 float), categorical (10 variables

1 - age (numeric)

2 - job : type of job (categorical: 'admin.','blue collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')

3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)

4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')

5 - default: has credit in default? (categorical: 'no','yes','unknown')

6 - housing: has housing loan? (categorical: 'no','yes','unknown')

7 - loan: has personal loan? (categorical: 'no','yes','unknown')

# related with the last contact of the current campaign:

8 - contact: contact communication type (categorical: 'cellular','telephone')

9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

10 - day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')

11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

# other attributes:

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

14 - previous: number of contacts performed before this campaign and for this client (numeric)

15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')

# social and economic context attributes

16 - emp.var.rate: employment variation rate - quarterly indicator (numeric)

17 - cons.price.idx: consumer price index - monthly indicator (numeric)

18 - cons.conf.idx: consumer confidence index - monthly indicator (numeric)

19 - euribor3m: euribor 3 month rate - daily indicator (numeric)

20 - nr.employed: number of employees - quarterly indicator (numeric)

\*Label (output) type: binary categorical

**\*If Label Type is Categorical, is the number of samples significantly unbalanced (maximal variation of more than a factor of 2)**? Yes

**If yes, \*rate as:**

8, significant (maximal variation is factor of 2 to 10)

**\*Has Missing Data**? No.

**\*Is the problem/dataset a Kaggle competition (current or past)?** No

**If yes, answer:**

**\*(i) Is the competition current (give the end date), or past**?

Not Kaggle Competition

**\*(ii) How much information is available on the Kaggle website (e.g., in “kernels” and links therein)**? Not Kaggle Competition

**Any other comments on the dataset:** None