MATH 11205: Machine Learning in Python 2023-2024
Project 2 Description

We will be using a subset of the data collected by Freddie Mac, called the Single Family Loan-Level Dataset. A simplified version of the data will be used for this project, provided in the file `freddiemac.csv`, after some initial cleaning steps (described below). This smaller data set combines and focuses only the years of 2017 and 2018. Only loans that have defaulted or been paid off (no active loans) are considered. But it is highly imbalanced, with only 113/6104= 1.85% of defaults. Make sure that, you deal with the imbalanced data issue during your data pre-processing.

Assignment Goal

For the purpose of the project, consider yourself a **Data Scientist Consultant** who has been hired by Freddie Mac to analyse loan-level credit performance data on fully amortizing fixed-rate Single-Family mortgages purchased or guaranteed by the company. Generally, the goal is to use the data to help investors build more accurate credit performance models in support of ongoing risk sharing initiatives highlighted by regulators, specifically, the Federal Housing Finance Agency. For further details about the data, please see the main webpage for the Single Family Loan-Level Dataset.

Towards this aim, you have been asked to use this data to build a classification model to predict if a client will default. In particular, the company is interested in identifying important factors that have impact on defaulting. In summary, you need to develop an **explainable, validated** classification model for **default** as the binary outcome of interest using features derived from the data provided and any additional sources you would like to use. Modeling-wise, you can think of various classification options that we covered after the flexible learning week, if the data set and the problem nature fits to their implementation.

It is important that any conclusions you draw from your model are well supported and sound and that you understand limitations of the model and the data. We explicitly **do not want a blackbox model** - you should be able to explain and justify your modeling choices and your model's predictions.

Your model may use as few or as many of the provided variables, and you may transform and manipulate these variables in any way that you want to generate additional features.

We have covered a number of models and modeling approaches in the lectures and workshops, and you should explore a variety of different approaches for this particular task. However, your ultimate goal is to deliver a **single** model. These are competing interests, and it is up to you to find a reasonable balance between exploring different models and selecting your proposed model; some of your marks will be based on how well you accomplish this. In addition, you should compare the performance of your model against a baseline model(s); although the main focus should be the description of your model (not the baseline).

Working as a team

This project may be completed by a team of up to 4 students (minimum of 1 student). Feel free to create your own team during workshop hours, building on the pairs for the workshop assignments. Since we are not assigning teams, if you are a team that is looking for more members or someone looking for a team please use the pinned post on Piazza to find each other.

After the assignment is completed we will distribute a brief peer evaluation survey - members who contributed significantly less than their peers will potentially have their overall mark penalized.

Dataset Details

These are the available variables given in the dataset `freddiemac.csv`:

- `fico` - CREDIT SCORE: a number, prepared by third parties, summarizing the borrower's creditworthiness, which may be indicative of the likelihood that the borrower will timely repay future obligations. Generally, the credit score disclosed is the score known at the time of acquisition and is the score used to originate the mortgage. Numeric with values between $300 - 850$ and 9999 for not available (credit scores < 300 or > 850 are be shown as not available).
- `dt_first_pi` - FIRST PAYMENT DATE: the date of the first scheduled mortgage payment due under the terms of the mortgage note. Format YYYYMM.
- `flag_fthb` - FIRST TIME HOMEBUYER FLAG: indicates whether the Borrower, or one of a group of Borrowers, is an individual who (1) is purchasing the mortgaged property, (2) will reside in the mortgaged property as a primary residence, and (3) had no ownership interest (sole or joint) in a residential property during the three-year period preceding the date of the purchase of the mortgaged property. With certain limited exceptions, a displaced homemaker or single parent may also be considered a First-Time Homebuyer if the individual had no ownership interest in a residential property during the preceding three-year period other than an ownership interest in the marital residence with a spouse. Format: Y=Yes, N=No, 9 = Not Available or Not Applicable.
- `dt_matr` - MATURITY DATE: the month in which the final monthly payment on the mortgage is scheduled to be made as stated on the original mortgage note. Format YYYYMM.
- `cd_msa` - METROPOLITAN STATISTICAL AREA (MSA) OR METROPOLITAN DIVISION: code, with null indicating that the area in which the mortgaged property is located is a) neither an MSA nor a Metropolitan Division, or b) unknown.
- `mi_pct` - MORTGAGE INSURANCE PERCENTAGE (MI %): the percentage of loss coverage on the loan, at the time of Freddie Mac's purchase of the mortgage loan that a mortgage insurer is providing to cover losses incurred as a result of a default on the loan. Amounts of mortgage insurance reported by Sellers that are less than 1% or greater than 55% will be disclosed as "Not Available," which will be indicated 999. No MI will be indicated by zero.
- `cnt_units` - NUMBER OF UNITS: denotes whether the mortgage is a one-, two-, three-, or four-unit property, with 99 indicating Not Available.
- `occpy_sts` - Denotes whether the mortgage type is owner occupied (P), second home (S), or investment property (I), or not available (9).
- `cltv` - ORIGINAL COMBINED LOAN-TO-VALUE (CLTV): with 999 indicating not available.

- `dti` - ORIGINAL DEBT-TO-INCOME (DTI) RATIO: disclosure of the debt to income ratio is based on (1) the sum of the borrower's monthly debt payments, including monthly housing expenses that incorporate the mortgage payment the borrower is making at the time of the delivery of the mortgage loan to Freddie Mac, divided by (2) the total monthly income used to underwrite the loan as of the date of the origination of the such loan. Ratios greater than 65% are indicated that data is Not Available (999).
- `orig_upb` - The UPB of the mortgage on the note date (rounded to the nearest $1,000).
- `ltv` - ORIGINAL LOAN-TO-VALUE (LTV). Values: 6% - 105%, with 999 = Not Available
- `int_rt` - The interest rate of the loan as stated on the note at the time the loan was originated.
- `channel` - indicates whether a Broker or Correspondent, originated or was involved in the origination of the mortgage loan. If a Third Party Origination is applicable, but the Seller does not specify Broker or Correspondent, the disclosure will indicate "TPO Not Specified". Similarly, if neither Third Party Origination nor Retail designations are available, the disclosure will indicate "TPO Not Specified." If a Broker, Correspondent or Third Party Origination disclosure is not applicable, the mortgage loan will be designated as Retail. Values: R = Retail, B = Broker, C = Correspondent, T = TPO Not Specified, 9 = Not Available.
- `ppmt_pnlty` - PREPAYMENT PENALTY MORTGAGE (PPM) FLAG: denotes whether the mortgage is a PPM. A PPM is a mortgage with respect to which the borrower is, or at any time has been, obligated to pay a penalty in the event of certain repayments of principal.
- `prod_type` - Denotes that the product is a fixed-rate mortgage or adjustable-rate mortgage.
- `st` - A two-letter abbreviation indicating the state or territory within which the property securing the mortgage is located.
- `prop_type` - Denotes whether the property type secured by the mortgage is a condominium (CO), planned unit development (PU), cooperative share (CP), manufactured home (MH), or Single-Family home (SF). If the Property Type is Not Available, this will be indicated by 99.
- `zipcode` - The postal code for the location of the mortgaged property. Format ###00, where ### represents the first three digits of the 5- digit postal code and 00 = Unknown.
- `id_loan` - Unique identifier assigned to each loan
- `loan_purpose` - Indicates whether the mortgage loan is a Cash-out Refinance mortgage (C), No Cash-out Refinance mortgage (N), Refinance mortgage not specified (R), or a Purchase mortgage (P), with 9 =Not Available.
- `orig_loan_term` - ORIGINAL LOAN TERM: the number of scheduled monthly payments of the mortgage based on the First Payment Date and Maturity Date.
- `cnt_borr` - The number of Borrower(s) who are obligated to repay the mortgage note secured by the mortgaged property. Disclosure denotes only whether there is one borrower (1), or more than one borrower associated with the mortgage note (2).
- `seller_name` - SELLER NAME: the entity acting in its capacity as a seller of mortgages to Freddie Mac at the time of acquisition.
- `servicer_name` - SERVICER NAME: the entity acting in its capacity as the servicer of mortgages to Freddie Mac as of the last period for which loan activity is reported in the Dataset.
- `flag_sc` - SUPER CONFORMING FLAG: For mortgages that exceed conforming loan limits with origination dates on or after 10/1/2008 and were delivered to Freddie Mac on or after 1/1/2009.
- `default` - whether person defaulted or not, binary response

You may choose to utilise additional data sources, for example, when encoding categorical features, if you feel that it is useful. In this case, please describe and reference any additional sources.

Required Structure

A Jupyter notebook template called 'project2.ipynb' has been provided. It includes the required sections along with brief instructions on what should be included in each section. Your completed assignment must follow this structure - **you should not add or remove any of these sections, if you feel it is necessary you may add extra subsections within each**. Please remove the instructions for each section in the final document.

All of your work must be contained in the 'project2.ipynb' notebook, we will only mark what is included in this file (both the write-up and relevant coding). You may work on the notebook in whichever environment you prefer, but please ensure that the final pdf file includes all necessary parts of your writing.

Our expectation is that most projects will be roughly 20-25 pages in length at most including text & figures, but excluding the related code. Overall, there is an **upper limit of 30 pages** including the coding part. Your notebook must include all of your work, but make sure that you are only retaining required components, e.g. remove unused code and figures (if a figure is not explicitly discussed in the text it should not be in the final document). **So, there is a trade-off between the length of your text and coding snippets while constructing your report.** Overall, your project will be partially assessed on your organization / presentation of the document - it should be as polished and streamlined as possible. **Try to be as concise as possible while creating your write-up. We highly recommend that you check the appearance of your rendered PDF before submitting, as its appearance can differ significantly from the notebook**.

You are expected to submit your completed work. For this, please submit your final PDF of project report (generated from a Jupyter notebook) to the Project assignment on Gradescope. Please ensure that you **tag all groups members** on Gradescope, and also add all group member names either in the notebook metadata or in additional markdown cell block at the beginning of the file.

Getting Help

- **Week 11 Workshop:** There will be no notebook during Week 11. Instead we will focus on answering any project related questions.
- **Piazza:** This forum will be used as the central location for all course related discussions and questions, and should be used over emailing course staff directly. The course lecturers will monitor and respond to questions, but feel free to provide some constructive responses to peer's questions. You can access Piazza from the course LEARN page or sign-up at:

  https://piazza.com/ed.ac.uk/spring2023/math11205

  Also, see the good practice guide for how to use piazza most effectively:

  https://teaching.maths.ed.ac.uk/main/undergraduate/studies/learning-advice/piazza
- You can also ask questions at the end of lectures during any Q&A time or during workshops.

Further References

We have provided additional resources in the project materials.
For further information on the dataset and variables, see:

- *Single Family Loan-Level Dataset*:
  https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset

For further information, on predictors of default:

- Bhattacharya et al (2019) *A Bayesian approach to modeling mortgage default and prepayment.* European Journal of Operational Research, 274: 1112-1124.
  https://www.sciencedirect.com/science/article/pii/S0377221718309159