Final Case Study: ML-AI, Modelling

Introduction

In this case study you will help an organization which funds projects submitted by high school teachers across US. The name of the organization is Donor's Choose. DonorsChoose.org is an online charity that makes it easy to help students in need through school donations. At any time, thousands of teachers in K-12 schools propose projects requesting materials to enhance the education of their students. When a project reaches its funding goal, they ship the materials to the school.

In this exercise the goal is to identify projects that are exceptionally exciting to the business, at the time of posting. While all projects on the site fulfill some kind of need, certain projects have a quality above and beyond what is typical. By identifying and recommending such projects early, Donor's Choose will improve funding outcomes, better the user experience, and help more students receive the materials they need to learn.

In order to complete this project, you will need to use a broad range of skills in text processing, feature engineering and predictive modelling.

Data

The data for this case study is provided in the form of csv files. There are in total 3 files:

1. essays.csv: contains project text posted by teachers.
2. projects.csv: contains information about each project
3. outcomes.csv: contains information about the outcomes of projects

Any project posted before 2014-01-01 is in the training set (along with its funding outcomes). Any project posted after that is in the test set.

Data Dictionary

Below is a brief explanation of the provided data fields. Descriptions of self-explanatory names are omitted.

outcomes.csv

is_exciting - ground truth of whether a project is exciting from business perspective (target variable, model for exciting projects)

at_least_1_teacher_referred_donor - teacher referred = donor donated because teacher shared a link or publicized their page

fully_funded - project was successfully completed

at_least_1_green_donation - a green donation is a donation made with credit card, PayPal, Amazon or check

great_chat - project has a comment thread with greater than average unique comments

three_or_more_non_teacher_referred_donors - non-teacher referred is a donor that landed on the site by means other than a teacher referral link/page

one_non_teacher_referred_donor_giving_100_plus - see above

donation_from_thoughtful_donor - a curated list of ~15 donors that are power donors and picky choosers (we trust them selecting great projects)

great_messages_proportion - how great_chat is calculated. proportion of comments on the project page that are unique. If > avg (currently 62%) then great_chat = True

teacher_referred_count - number of donors that were teacher referred (see above)

non_teacher_referred_count - number of donors that were non-teacher referred (see above)

projects.csv

projectid - project's unique identifier

teacher_acctid - teacher's unique identifier (teacher that created a project)

schoolid - school's unique identifier (school where teacher works)

school_ncesid - public National Center for Ed Statistics id

school_latitude

school_longitude

school_city

school_state

school_zip

school_metro

school_district

school_county

school_charter - whether a public charter school or not (no private schools in the dataset)

school_magnet - whether a public magnet school or not

school_year_round - whether a public year round school or not

school_nlns - whether a public nlns school or not

school_kipp - whether a public kipp school or not

school_charter_ready_promise - whether a public ready promise school or not

teacher_prefix - teacher's gender

teacher_teach_for_america - Teach for America or not

teacher_ny_teaching_fellow - New York teaching fellow or not

primary_focus_subject - main subject for which project materials are intended

primary_focus_area - main subject area for which project materials are intended

secondary_focus_subject - secondary subject

secondary_focus_area - secondary subject area

resource_type - main type of resources requested by a project

poverty_level - school's poverty level.

highest: 65%+ free of reduced lunch

high: 40-64%

moderate: 10-39%

low: 0-9%

grade_level - grade level for which project materials are intended

fulfillment_labor_materials - cost of fulfillment

total_price_excluding_optional_support - project cost excluding optional tip that donors give to DonorsChoose.org while funding a project

total_price_including_optional_support - project cost including optional tip that donors give to DonorsChoose.org while funding a project

students_reached - number of students impacted by a project (if funded)

eligible_double_your_impact_match - project was eligible for a 50% off offer by a corporate partner (logo appears on a project, like Starbucks or Disney)

eligible_almost_home_match - project was eligible for a $100 boost offer by a corporate partner

date_posted - date a project went live on the site

essays.csv

projectid - unique project identifier

teacher_acctid - teacher id that created a project
title - title of the project
short_description - description of a project
need_statement - need statement of a project
essay - complete project essay

Links to data

outcomes.csv: https://s3.us-east-2.amazonaws.com/datafaculty/final_case/outcomes.csv.zip
projects.csv: https://s3.us-east-2.amazonaws.com/datafaculty/final_case/projects.csv.zip
essays.csv: https://s3.us-east-2.amazonaws.com/datafaculty/final_case/essays.csv.zip
sample_data_audit.csv: https://s3.us-east-2.amazonaws.com/datafaculty/final_case/sample_data_audit.csv

Deliverables

Following are the submission requirements for this case study:

1. Data Audit report and code used to create the data audit report. You can find a sample data audit report in the data folder; the name of the file is sample_data_audit.csv. You will need to submit data audit report for each data set along with the code.
2. Feature engineering code in a jupyter notebook format. Make sure the code is properly commented out.
3. You will also need to submit the code for the final model selected by you, the model should be built to predict *if a project is exciting*.
4. AUC reported on five-fold CV done by you, using the final model selected in 3. The AUC should be reported in the following format in a csv file:

Fold, AUC
1,Value
2, Value
3,Value
4,Value
5,Value

5. Probability predictions for test data using the model finalized in 3. The predictions should also be submitted as a csv file with the following format:

is_exciting, projectid
0.08, projectid1
0.9, projectid2
(Note project ids are actual project ids, once you separate the test data, this will be alphanumeric string in the actual file)

Hints and Guidelines:

1. This project has data which is both structured and unstructured (text)
2. You will need to figure out a way to join the three files appropriately, so that you can proceed with the feature engineering.
3. If you are creating a count matrix or tfidf matrix while extracting text features, make sure you put a limit to the size of the vocabulary. Since the data is relatively big, count matrix or tfidf matrix will take a very long time to build. For example if the

vocabulary size is 1000, it takes around 10 minutes to build the tfidf matrix on a machine with quad core processor and 8 gb of RAM

4. Make sure that you remove numbers and other unnecessary tokens from the text before you create a tfidf or count matrix.
5. Make sure that you separate test data and train data before you start feature engineering.
6. Use the data audit reports, to eliminate variables that will not be used for modelling.
7. Start with simple classifiers such as a logistic regression before you try ensemble models.
8. Keep in mind that ensemble models will consume a lot of computing time, for example a Random Forest model with 1000 trees using text features takes one hour on a quad core cpu with 8gb of RAM.
9. To do 5 fold CV,use the kfold module in model_selection() module of sklearn
10. You will submit your solutions on the JLC, the solutions aren't supposed to be emailed.
11. Since this is a final case study, the faculty support will be limited to only administrative queries, we will not review your in-progress code or syntax related queries. You are free to discuss queries related to approach to the problem, though.

On a side note since the data set size is relatively large, it is recommended that you work on a system with atleast 8GB of RAM. If your system RAM is less than 8GB, then you will need to use sample data to build the model.