CS 1675
Introduction to Machine Learning

Final project

Test Set

Reporting to PPG for bonus

A hold out test set of input values is provided on Canvas

	A	B	C	D	E	F	G	H	I	J
1	x1	x2	x3	x4	v1	v2	v3	v4	v5	m
2	0.107736	0.022262	0.113261	0.050347	6.423952	0.205828	1.892627	0.105402	0.671098	A
3	0.300597	0.05864	0.106367	0.049486	6.211491	0.392679	2.120956	0.114334	2.107955	A
4	0.496829	0.03117	0.093909	0.049447	1.383901	0.800464	7.940589	0.43957	9.841814	A
5	0.244908	0.0339	0.210922	0.048939	8.891101	0.800688	1.931002	0.451399	9.175254	A
6	0.38847	0.053426	0.197543	0.049832	3.868564	0.206151	8.039055	0.586806	6.34117	A
7	0.110522	0.051354	0.269977	0.051235	1.34079	0.555273	2.061072	0.565206	9.268104	A
8	0.492397	0.033276	0.239787	0.050228	6.287046	0.605271	8.040479	0.887996	4.91895	A
9	0.255758	0.055675	0.328047	0.049999	1.272192	0.409905	7.724796	0.224859	8.193524	A
10	0.397577	0.051946	0.322589	0.050012	6.383613	0.158705	8.110438	0.922231	8.779091	A
11	0.116364	0.05887	0.380113	0.05218	8.932519	0.208629	2.102075	0.210848	9.081519	A
12	0.297918	0.055397	0.409473	0.049827	8.735747	0.603104	1.739974	0.858853	9.992252	A
13	0.091201	0.150673	0.103102	0.051092	1.316315	0.188295	7.875094	0.685057	2.382675	A
14	0.294276	0.148168	0.102609	0.049967	1.200748	0.583337	7.964343	0.613	2.169496	A
15	0.499877	0.161462	0.112433	0.051802	3.639316	0.620095	8.198351	0.306016	2.076829	A
16	0.235662	0.162807	0.203541	0.050368	8.851131	0.584698	8.017373	0.77144	0.524538	A
17	0.390153	0.150147	0.204386	0.051173	6.102208	0.834874	7.950265	0.973788	5.503631	A

The input names are the same as those in the training set

You must predict the continuous response and the binary outcome using this test set

- You must select 1 regression model and 1 classification model.
- You must predict the continuous output.
- You must predict the probability of the event.
- You must classify the binary outcome assuming a default threshold of 0.5.

Organize the test set predictions

- Compile the predictions into a dataframe with 4 columns:
  - `id` – the row index (use `tibble::rowid_to_column()` function)
  - `y` – the prediction for the logit-transformed continuous response
  - `outcome` – the classified outcome
    - Must have values `event` and `non_event`
  - `probability` – the predicted probability of the event

- Save the dataframe to a CSV file.
  - Can save using the `readr::write_csv()` function

Example CSV file of the predictions

- I created some example models (maybe not that great), and saved the predictions from those models accordingly.

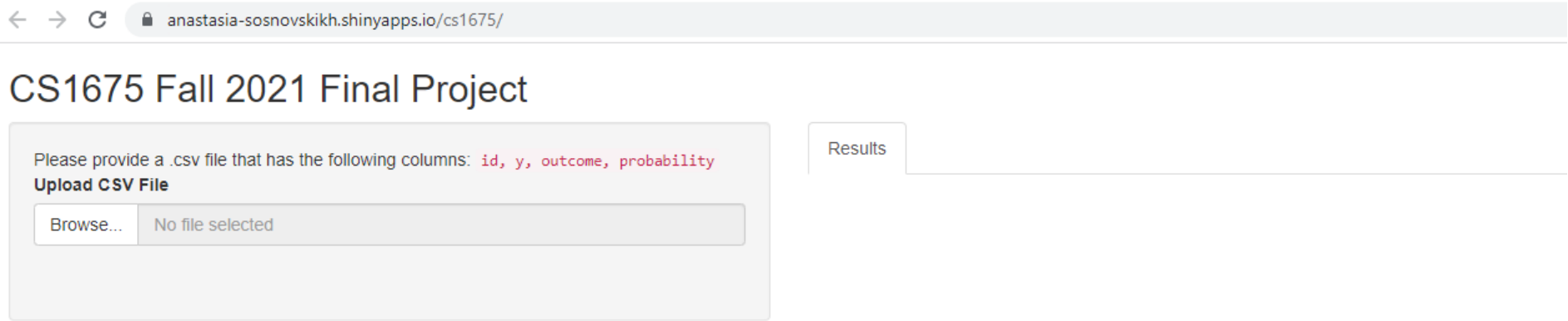
- The CSV file MUST have these 4 columns with these exact names.

	A	B	C	D
1	id	y	outcome	probability
2	1	1.40773	non_even	0.194182
3	2	-0.09687	non_even	0.260732
4	3	-1.25394	event	0.610336
5	4	-0.34606	event	0.532572
6	5	-1.49484	event	0.744715
7	6	0.74325	non_even	0.289444
8	7	-0.92859	event	0.596826
9	8	-1.12712	event	0.723652
10	9	-1.37852	event	0.753914
11	10	0.283609	non_even	0.398751

Your predictions will “scored” by uploading the predictions to a website

- Go to the following R Shiny app.

- The landing page looks like:



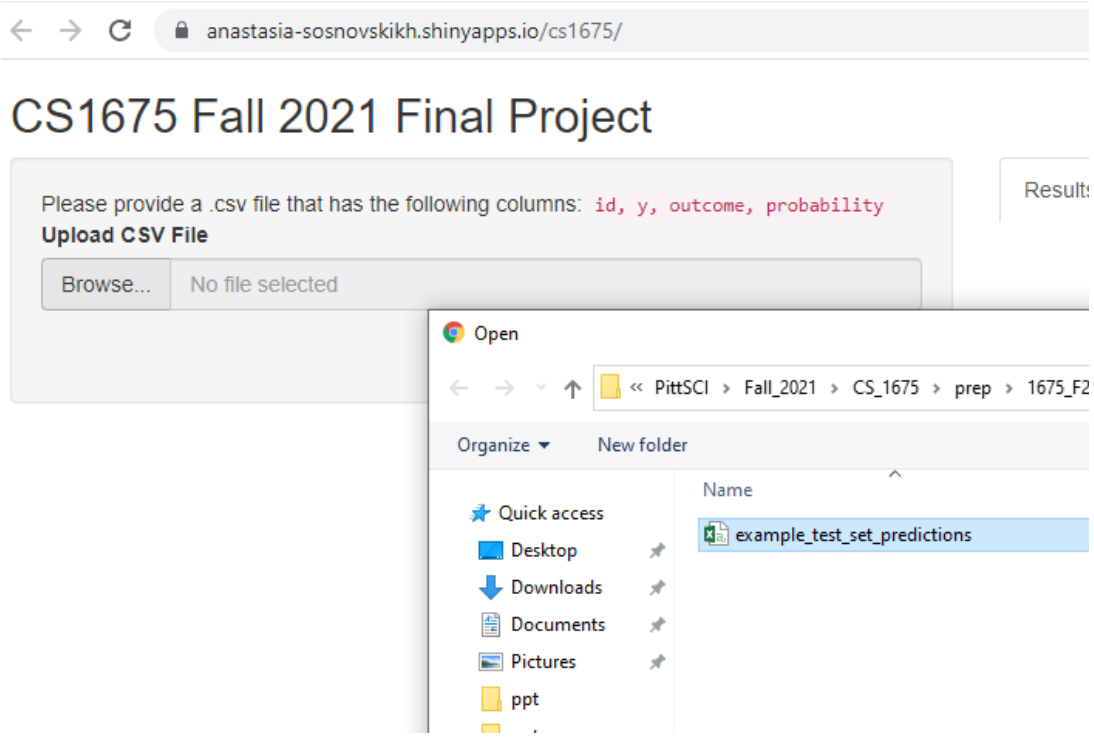
Your predictions will “scored” by uploading the predictions to a website

- Go to the following R Shiny app.

- Select the Browse button and upload your CSV file of predictions to the website.

- I named my example CSV file example_test_set_predictions.csv

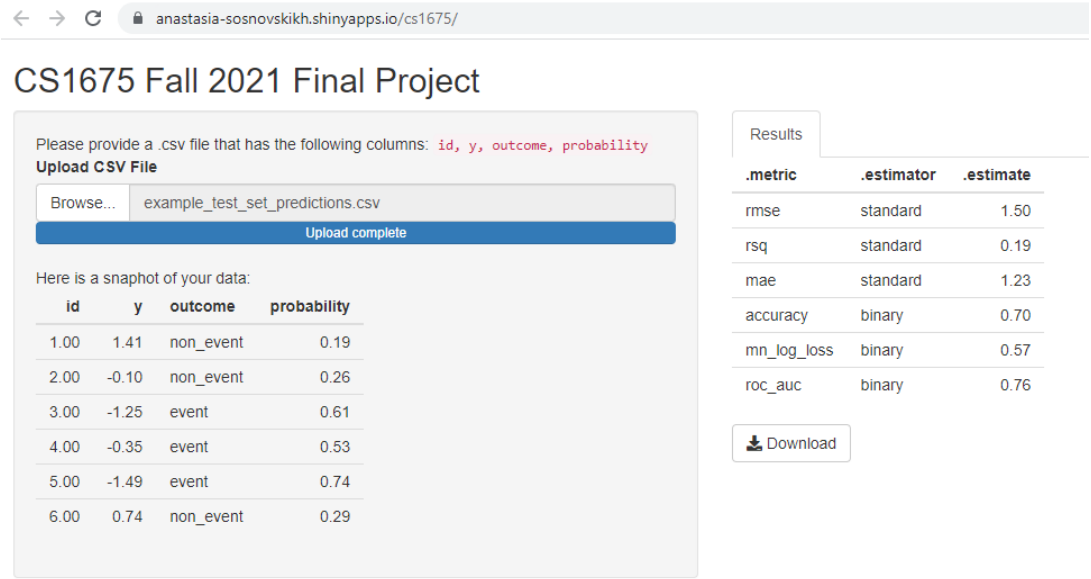
- You may name your CSV file whatever you want.



Your predictions will “scored” by uploading the predictions to a website

- Once uploaded the performance metrics on the hold-out test set will be shown to you.

- Press the Download button to save the performance metrics to your computer.



You <u>MUST</u> submit the downloaded CSV file as part of your final project submission

- The downloaded CSV file must be uploaded to Canvas along with all of your rendered HTML files and source .Rmd files.

- Do **<u>NOT</u>** zip files!!!!!!!!!!!! Upload each file.

<u>BONUS</u>: 7 points

- Create a short presentation which shows:
  - What are the most important inputs?
  - What are the trends of the logit-transformed response with respect to the most important inputs?
  - What are the trends of the probability of the event with respect to the most important inputs?
  - What input values do you recommend to minimize the fraction of corroded surface?

- Submit your presentation as a Power Point .pptx file.