ATMS 597 Project 5

Group G | Location: KSTC

Arka Mitra, Sarah Szymborski, and Jeff Thayer

DATA PROCESSING

Data Cleanup #1 - Metar Files Dataframe

- Wrote code to sift through disordered text in the metar files.
- Variable columns, variable string lengths, additional optional remarks and unaccounted UTF-8 reading errors had to be accounted for.
- Code reads line-by-line and looks for identifier strings, such as wind-data ends in ‘KT’, if at all. Reads them into dataframe.

	StnNo	ContainsDate	Time	Interval	Sttn	UTC	Type	Wind	Weath/Obstr	SkyCond	Temp	n_days	MSLP	RelHum		WndDir/Spd	WindMagnetic	RMK	AO2	X	X	X
0	14926KSTC	STC20000101000010001/01/00	00:00:32	5-MIN	KSTC	010600Z	AUTO	32010KT	10SM	CLR	00/M06	A2977	1170	63	-300	320/10	RMK	AO2	NaN	NaN	NaN	NaN
1	14926KSTC	STC20000101001010801/01/00	00:10:31	5-MIN	KSTC	010610Z	AUTO	31012KT	10SM	CLR	M01/M06	A2978	1150	66	-400	310/12	RMK	AO2	PRESRR	NaN	NaN	NaN
2	14926KSTC	STC20000101002010101/01/00	00:20:31	5-MIN	KSTC	010620Z	AUTO	31012KT	10SM	CLR	M01/M07	A2979	1150	66	-500	310/12	RMK	AO2	NaN	NaN	NaN	NaN
3	14926KSTC	STC20000101003010101/01/00	00:30:31	5-MIN	KSTC	010630Z	AUTO	32013KT	10SM	CLR	M01/M06	A2979	1150	66	-400	320/13	RMK	AO2	NaN	NaN	NaN	NaN
4	14926KSTC	STC20000101004010101/01/00	00:40:31	5-MIN	KSTC	010640Z	AUTO	32011KT	10SM	CLR	M01/M06	A2979	1140	69	-500	310/11	RMK	AO2	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4159	14926KSTC	STC20000131231510401/31/00	23:15:31	5-MIN	KSTC	010515Z	AUTO	35010KT	10SM	OVC025	M03/M08	A3016	810	71	-1200	350/10	RMK	AO2	NaN	NaN	NaN	NaN
4160	14926KSTC	STC20000131232510401/31/00	23:25:31	5-MIN	KSTC	010525Z	AUTO	35010KT	10SM	OVC023	M03/M08	A3016	810	71	-1200	350/10	RMK	AO2	NaN	NaN	NaN	NaN
4161	14926KSTC	STC20000131233510401/31/00	23:35:31	5-MIN	KSTC	010535Z	AUTO	35010KT	10SM	OVC023	M03/M08	A3016	800	71	-1200	350/10	RMK	AO2	NaN	NaN	NaN	NaN
4162	14926KSTC	STC20000131234510401/31/00	23:45:31	5-MIN	KSTC	010545Z	AUTO	35009KT	10SM	OVC023	M03/M08	A3016	800	71	-1200	340/09	RMK	AO2	NaN	NaN	NaN	NaN
4163	14926KSTC	STC20000131235510401/31/00	23:55:31	5-MIN	KSTC	010555Z	AUTO	35009KT	10SM	OVC023	M03/M08	A3017	800	71	-1200	350/09	RMK	AO2	NaN	NaN	NaN	NaN

DATA CLEANUP #2 - Input Variables Dataframe

1) Date and Time of Day (as a fraction of 24 hours)
2) Winds (Speed/Direction, Gusts if any and Variable winds, if any)
3) Temperature, Relative Humidity, dew point temperature
4) Precipitation accumulations (if any) over the last hour, 6 hours and 24 hours (in inches)
5) Precipitation type into 2 classes (Solid & Liquid)

	date	timeofday	Wind	Wind_Direction	Wind_Speed	Gusts	Variable_Winds	Temperature	Dewpoint	rhum	hourly	sixhourly	dayprecip	solid	liquid
0	2000-06-01 00:00:32	0.000370	1	40	10	0	0	17.0	6.0	49.0	NaN	NaN	NaN	0	0
1	2000-06-01 00:10:31	0.007303	1	50	12	1	0	17.0	6.0	49.0	NaN	NaN	NaN	0	0
2	2000-06-01 00:20:31	0.014248	1	50	9	0	0	16.0	6.0	51.0	NaN	NaN	NaN	0	0
3	2000-06-01 00:30:31	0.021192	1	50	9	1	0	16.0	6.0	NaN	NaN	NaN	NaN	0	0
4	2000-06-01 00:40:31	0.028137	1	30	10	1	0	16.0	6.0	51.0	NaN	NaN	NaN	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3333	2000-06-30 23:10:31	0.965637	1	190	6	0	0	18.0	14.0	75.0	NaN	NaN	NaN	0	0

DATA CLEANUP #3 - A few last minute fixes

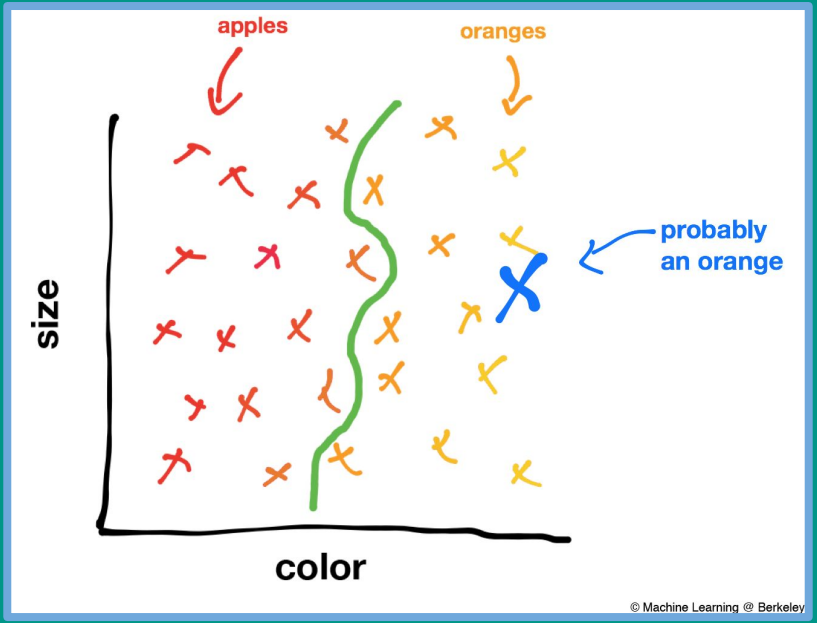
1) Fix temperatures and dew point temperatures that were flipped
2) Add another column, **prcp_type**, to store NaN, 0, and 1 for non-precipitating days, liquid, and frozen respectively
3) Convert date to day of year
4) Change negative RHUM values to NaN
5) Filter out all NaNs

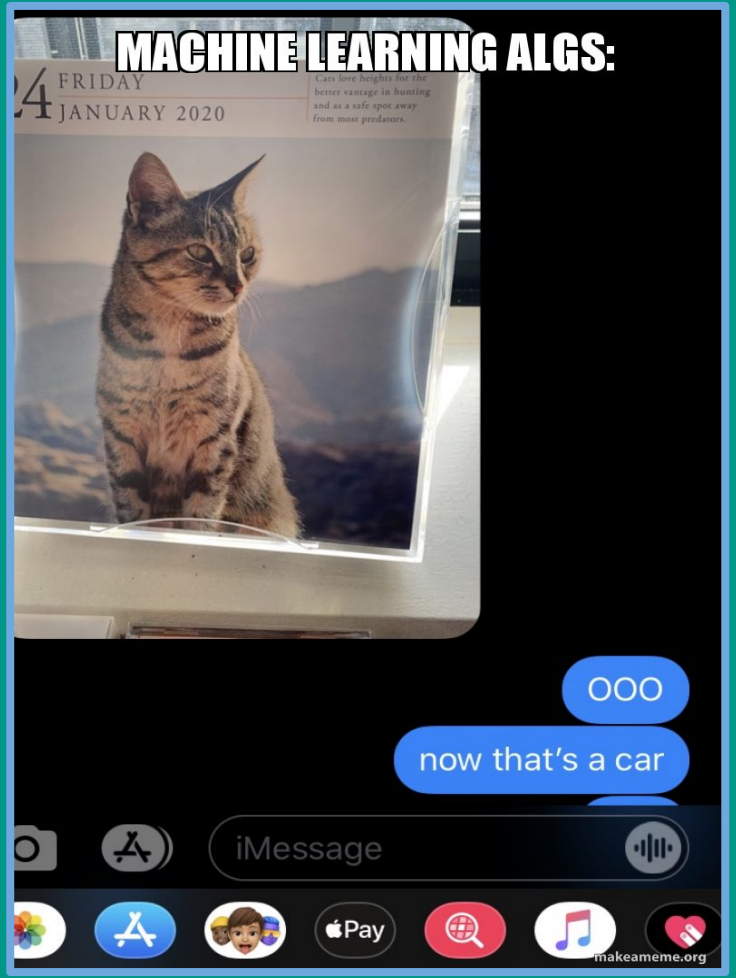
Final Sample Size: 6231 Obs

	date	timeofday	Temperature	Dewpoint	rhum	prcp_type
60960	080	0.868414	3.0	3.0	96.0	0.0
60961	080	0.875359	3.0	2.0	96.0	0.0
60962	080	0.882303	2.0	2.0	100.0	0.0
60963	080	0.889248	2.0	2.0	100.0	0.0
60964	080	0.896192	1.0	1.0	100.0	0.0
...	...	...	...	...	...	...
981232	346	0.375359	16.0	13.0	80.0	1.0
981237	346	0.410081	15.0	12.0	80.0	1.0
983181	360	0.847581	7.0	4.0	81.0	1.0
983447	364	0.566331	7.0	4.0	81.0	1.0
983448	364	0.573275	7.0	4.0	81.0	1.0

[6231 rows x 6 columns]

Results


apples
oranges
probably
an orange
size
color
© Machine Learning @ Berkeley


MACHINE LEARNING ALGS:
FRIDAY
JANUARY 2020
OOO
now that's a car
iMessage
Pay
makeameme.org

We scored 3 different ways: Model Score, Brier Score, and Brier Skill Score.

Logistic Regression

Hyperparameters: max_iter = 10E10, random_state = 42

Training Scores

LR Score: 0.7984407246044485

Brier Score: 0.2015592753955515

Brier Skill Score: 0.45903779615265305

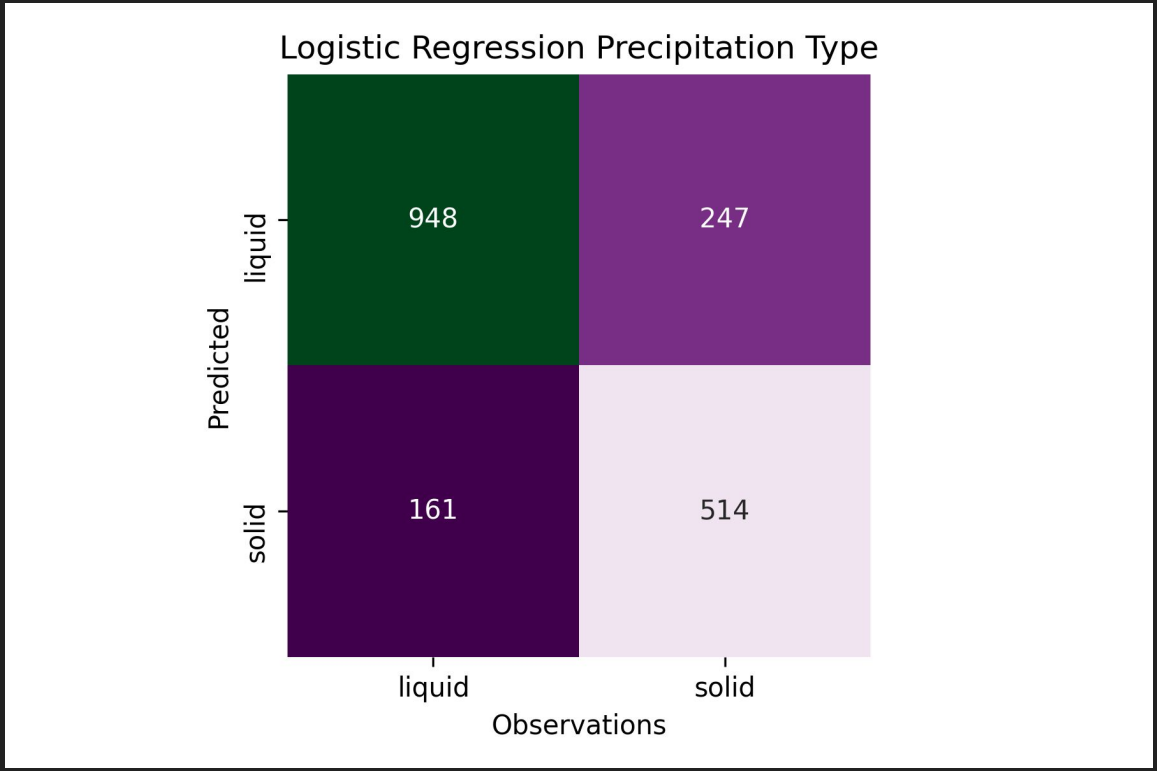
Testing Scores

LR Score: 0.7818181818181819

Brier Score: 0.21818181818181817

Brier Skill Score: 0.42380805369114083

Logistic Regression



Ensemble Random Forest

Hyperparameters: n_estimators = 100, random_state = 42
Also attempted: n_estimators = 100, random_state = 42, max_depth = 80, max_features = 'auto', min_samples_leaf = 1, min_samples_split = 3, bootstrap=True

Training Scores

LR Score: 1.0

Brier Score: 0.0

Brier Skill Score: 0.9921434622784686

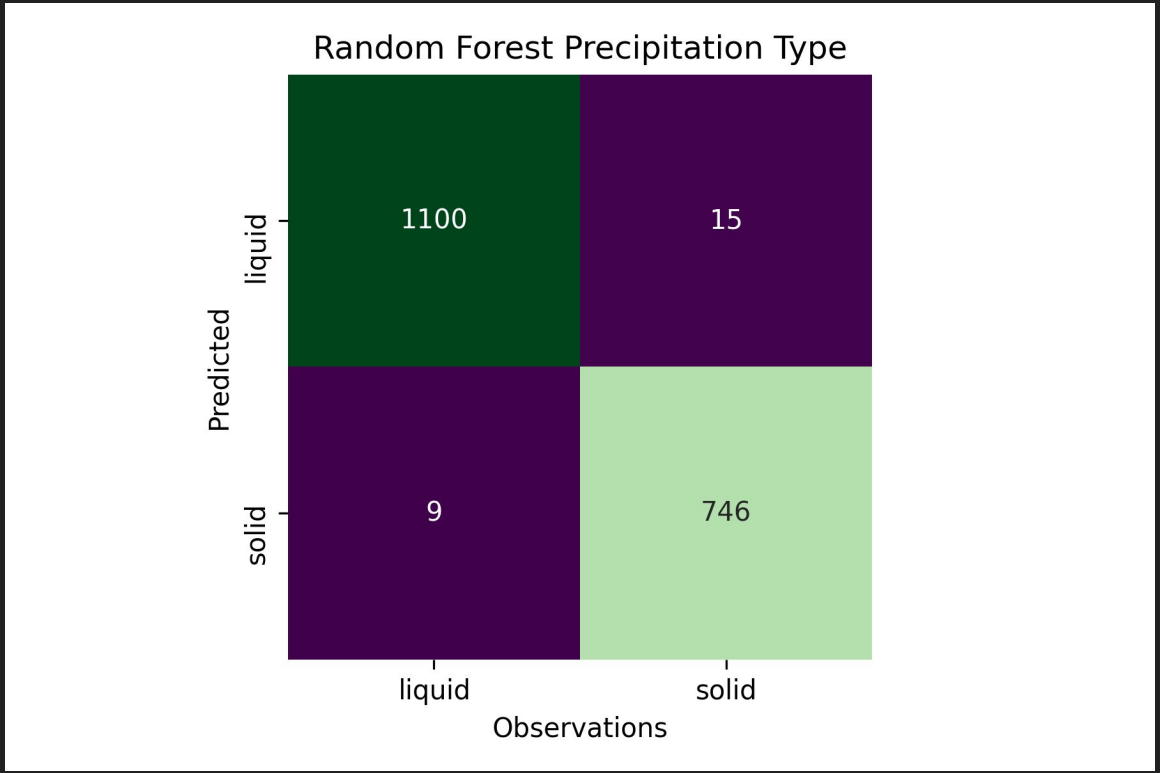
Testing Scores

LR Score: 0.9871657754010695

Brier Score: 0.012834224598930482

Brier Skill Score: 0.9495034036416893

Ensemble Random Forest



Improvements in Testing!

	Logistic Regression	Random Forest
Model Score	0.781	0.987
Brier Score	0.218	0.013
Brier Skill Score	0.424	0.949