Fundamentals of Machine Learning for Predictive Data Analytics

Chapter 10: Case Study - Galaxy Classification

John Kelleher and Brian Mac Namee and Aoife D'Arcy

john.d.kelleher@dit.ie brian.macnamee@ucd.ie aoife@theanalyticsstore.com

- The **Sloan Digital Sky Survey** (SDSS) is a landmark project that is cataloging the night sky in intricate detail and is facing exactly the problem described above.
- The SDSS telescopes collect over 175GB of data every night, and for the data collected to be fully exploited for science, each night sky object captured must be identified and cataloged within this data in almost real time.
- This case study describes the work undertaken when, in 2011, the SDSS hired Jocelyn, an analytics professional, to build a galaxy morphology classification model to include in their data processing pipeline.

Business Understanding

- The SDSS pipeline takes the data captured by the SDSS instruments and processes it, before storing the results of this processing in a centrally accessible database.
- The SDSS scientists wanted a system that could reliably classify galaxies into the important morphological (i.e., shape) types: **elliptical galaxies** and **spiral galaxies**.
- The scientists at SDSS wanted Jocelyn to build a machine learning model that could examine sky objects that their current rule-based system had flagged as being galaxies and categorize them as belonging to the appropriate morphological group.

(a) Elliptical (b) Clockwise spiral (c) Anti-clockwise spiral

Figure: Examples of the different galaxy morphology categories into which SDSS scientists categorize galaxy objects. (Credits for these images belong to the Sloan Digital Sky Survey, `www.sdss3.org`)

Data Understanding

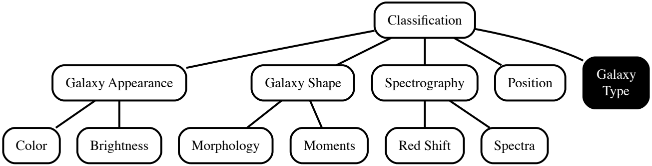
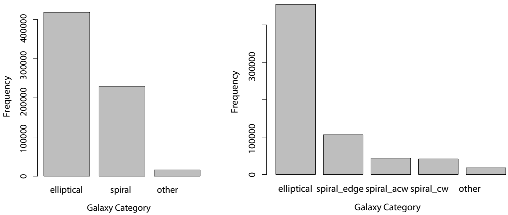


Figure: The first draft of the domain concepts diagram developed by Jocelyn for the galaxy classification task.

Name	Type	Description
objID	Continuous	Unique SDSS object identifier
p_el	Continuous	Fraction of votes for elliptical galaxy category
p_cw	Continuous	Fraction of votes for clockwise spiral galaxy category
p_acw	Continuous	Fraction of votes for anti-clockwise spiral galaxy category
p_edge	Continuous	Fraction of votes for edge-on disk galaxy category
p_mg	Continuous	Fraction of votes for merger category
p_dk	Continuous	Fraction of votes for don’t know category



(a) 3-level model (b) 5-level model

Figure: Bar plots of the different galaxy types present in the full SDSS dataset for the 3-level and 5-level target features.

Feature	Count	% Miss.	Card.	Min.	1^{st} Qrt.	Mean	Median	3^{rd} Qrt.	Max.	Std. Dev.
run	10 000	0.000	380	109.000	2 821.000	3 703.449	3 841.000	4 646.000	8 095.000	1 378.815
ra.1	10 000	0.000	9 964	0.032	151.376	185.258	185.015	220.555	359.990	59.116
dec.1	10 000	0.000	9 928	-11.234	9.707	24.867	23.414	39.107	69.826	18.919
rowc_u	10 000	0	1	0	0	0	0	0	0	0
rowc_g	10 000	0	1	0	0	0	0	0	0	0
rowc_r	10 000	0	1	0	0	0	0	0	0	0
rowc_i	10 000	0	1	0	0	0	0	0	0	0
rowc_z	10 000	0	1	0	0	0	0	0	0	0
skyIvar_u	10 000	0.000	9 986	-9 999.000	459.807	78.893	798.273	1 083.646	2 197.086	450.260
skyIvar_g	10 000	0.000	9 989	-9 999.000	439.550	965.879	2 957.923	6 005.711	9 913.587	2 766.697
skyIvar_r	10 000	0.000	9 988	-9 999.000	123.305	201.905	1 091.784	3 347.769	4 623.066	1 514.504
skyIvar_i	10 000	0.000	9 986	-9 999.000	46.019	174.790	434.484	1 825.934	2 527.567	851.422
skyIvar_z	10 000	0.000	9 986	-9 999.000	13.601	-234.234	49.569	75.388	205.066	44.511
psfMag_u	10 000	0.014	9 768	7.468	20.604	21.078	21.127	21.598	26.190	0.854
psfMag_g	10 000	0.014	9 743	8.299	19.057	19.479	19.539	19.967	26.169	0.778
psfMag_r	10 000	0.008	9 744	7.454	18.234	18.654	18.675	19.113	26.489	0.758
psfMag_i	10 000	0.008	9 744	7.332	17.833	18.274	18.263	18.722	25.456	0.804
psfMag_z	10 000	0.012	9 747	7.398	17.474	17.928	17.900	18.381	23.919	0.819
deVFlux_u	10 000	0.000	9 990	-3.683	11.643	43.053	23.074	44.313	28 616.040	194.727
deVFlux_g	10 000	0.000	9 987	-1 278.277	48.786	143.710	77.062	133.461	614 662.800	2 401.589
deVFlux_r	10 000	0.000	9 983	-4.368	111.038	267.736	152.745	250.646	137 413.000	993.654
deVFlux_i	10 000	0.000	9 980	-4.061	160.417	390.976	216.571	351.209	608 862.800	3 041.201
deVFlux_z	10 000	0.000	9 983	-14.720	204.723	528.685	276.991	447.445	2 264 700.000	9 073.949

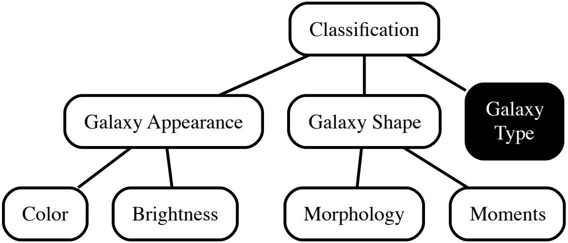


Figure: The revised domain concepts diagram for the galaxy classification task.

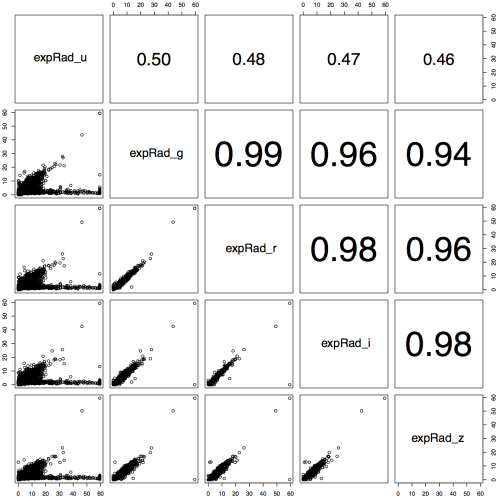


Figure: SPLOM diagrams of the EXPRAD measurement from the raw SDSS dataset. The SPLOM shows the measure across the five different photometric bands captured by the SDSS telescope (*u*, *g*, *r*, *i*, and *z*).

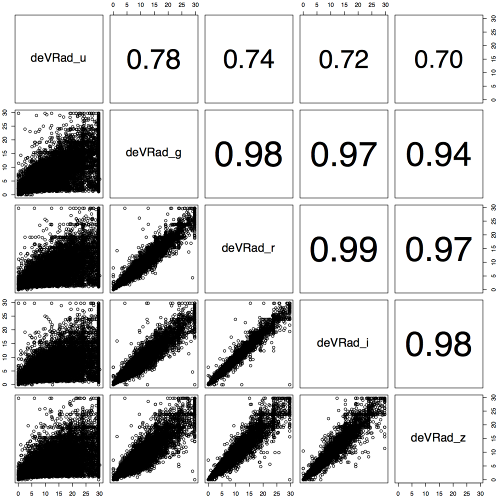
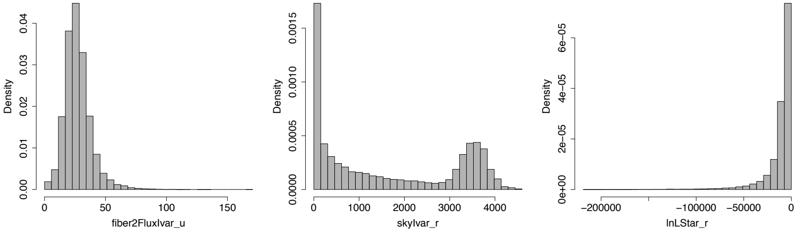


Figure: SPLOM diagrams of the DEVRAD measurement from the raw SDSS dataset. The SPLOM shows the measure across the five different photometric bands captured by the SDSS telescope (*u*, *g*, *r*, *i*, and *z*).

Data Preparation

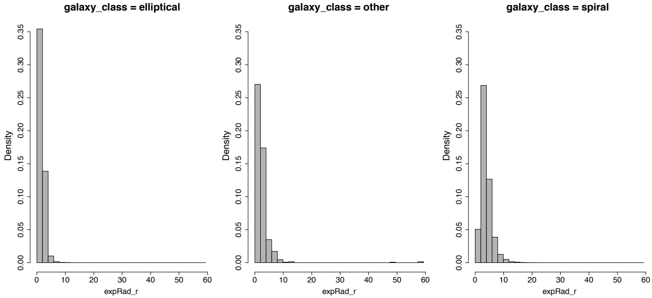
Feature	Feature	Feature
skyIvar_u/g/r/i/z	uErr_u/g/r/i/z	expFlux_u/g/r/i/z
psfMag_u/g/r/i/z	mE1_u/g/r/i/z	expFluxIvar_u/g/r/i/z
psfMagErr_u/g/r/i/z	mE2_u/g/r/i/z	modelFluxIvar_u/g/r/i/z
fiberMag_u/g/r/i/z	mE1E1Err_u/g/r/i/z	cModelFlux_u/g/r/i/z
fiberMagErr_u/g/r/i/z	mE1E2Err_u/g/r/i/z	cModelFluxIvar_u/g/r/i/z
fiber2Mag_u/g/r/i/z	mE2E2Err_u/g/r/i/z	aperFlux7_u/g/r/i/z
fiber2MagErr_u/g/r/i/z	mRrCc_u/g/r/i/z	aperFlux7Ivar_u/g/r/i/z
petroMag_u/g/r/i/z	mRrCcErr_u/g/r/i/z	lnLStar_u/g/r/i/z
petroMagErr_u/g/r/i/z	mCr4_u/g/r/i/z	lnLExp_u/g/r/i/z
psfFlux_u/g/r/i/z	deVRad_u/g/r/i/z	lnLDeV_u/g/r/i/z
psfFluxIvar_u/g/r/i/z	deVRadErr_u/g/r/i/z	fracDeV_u/g/r/i/z
fiberFlux_u/g/r/i/z	deVAB_u/g/r/i/z	dered_u/g/r/i/z
fiberFluxIvar_u/g/r/i/z	deVABErr_u/g/r/i/z	deredDiff_u_g
fiber2Flux_u/g/r/i/z	deVMag_u/g/r/i/z	deredDiff_g_r
fiber2FluxIvar_u/g/r/i/z	deVMagErr_u/g/r/i/z	deredDiff_r_i
petroFlux_u/g/r/i/z	deVFlux_u/g/r/i/z	deredDiff_i_z
petroFluxIvar_u/g/r/i/z	deVFluxIvar_u/g/r/i/z	petroRatio_i
petroRad_u/g/r/i/z	expRad_u/g/r/i/z	petroRatio_r
petroRadErr_u/g/r/i/z	expRadErr_u/g/r/i/z	aE_i
petroR50_u/g/r/i/z	expAB_u/g/r/i/z	petroMagDiff_u_g
petroR50Err_u/g/r/i/z	expABErr_u/g/r/i/z	petroMagDiff_g_r
petroR90_u/g/r/i/z	expMag_u/g/r/i/z	petroMagDiff_r_i
petroR90Err_u/g/r/i/z	expMagErr_u/g/r/i/z	petroMagDiff_i_z
q_u/g/r/i/z	cModelMag_u/g/r/i/z	galaxy_class_3
qErr_u/g/r/i/z	cModelMagErr_u/g/r/i/z	galaxy_class_5
u_u/g/r/i/z		

Feature	Count	% Miss.	Card.	Min.	1st Qrt.	Mean	Median	3rd Qrt.	Max.	Std. Dev.
skyIvar_u	640 432	0.000	639 983	0.000	465.525	784.780	793.201	1 079.525	2 190.047	447.360
skyIvar_g	640 432	0.000	640 081	0.000	442.549	3 318.724	2 949.622	6 008.313	9 898.472	2 769.840
skyIvar_r	640 432	0.000	640 178	0.000	127.179	1 629.862	1 094.925	3 342.651	4 596.461	1 513.383
skyIvar_i	640 432	0.000	640 042	0.000	48.284	842.175	436.128	1 825.877	2 515.348	852.733
skyIvar_z	640 432	0.000	640 042	0.000	13.896	52.194	49.763	75.098	205.685	44.194
mE2_g	640 432	0.000	629 246	-0.955	-0.134	0.008	0.010	0.151	0.969	0.280
fiber2FluxIvar_u	640 432	0.000	639 827	0.001	20.308	27.243	25.964	32.401	170.696	11.024
psfMag_u	640 432	0.000	632 604	13.757	20.591	21.052	21.117	21.577	25.564	0.810
petroFluxIvar_u	640 432	0.000	627 391	0.000	0.163	0.400	0.305	0.531	6.291	0.355
lnLStar_r	640 432	0.000	639 690	-218 875.300	-12 623.050	-12 009.952	-6 771.368	-4 308.989	0.000	16 193.728
petroMag_r	640 432	0.000	628 562	11.720	16.763	17.077	17.287	17.608	22.717	0.746
expAB_i	640 432	0.000	623 467	0.050	0.494	0.646	0.671	0.813	1.000	0.202
deredDiff_u_g	640 432	0.000	630 319	-2.474	1.291	1.608	1.665	1.892	6.674	0.395
deredDiff_g_r	640 432	0.000	631 627	-1.063	0.642	0.821	0.840	0.991	4.695	0.269
deredDiff_r_i	640 432	0.000	611 597	-4.464	0.355	0.391	0.403	0.444	2.221	0.100
deredDiff_i_z	640 432	0.000	615 131	-2.285	0.229	0.275	0.296	0.335	5.332	0.107
petroRatio_i	640 432	0.000	640 432	1.123	2.326	2.671	2.683	3.009	25.523	0.458
petroRatio_r	640 432	0.000	640 432	1.183	2.290	2.630	2.638	2.961	10.049	0.418
aE_i	640 432	0.000	640 432	0.000	0.125	0.269	0.226	0.378	0.903	0.183
modelMagDiff_u_g	640 432	0.000	630 476	-2.452	1.334	1.651	1.708	1.936	6.831	0.397
modelMagDiff_g_r	640 432	0.000	630 437	-1.049	0.675	0.854	0.873	1.025	4.748	0.270
modelMagDiff_r_i	640 432	0.000	613 667	-4.455	0.375	0.412	0.424	0.465	2.252	0.101
modelMagDiff_i_z	640 432	0.000	615 346	-2.271	0.248	0.294	0.315	0.354	5.340	0.107
petroMagDiff_g_r	640 432	0.000	631 901	-1.992	0.640	0.828	0.842	0.997	5.125	0.275
petroMagDiff_r_i	640 432	0.000	612 827	-3.322	0.353	0.392	0.406	0.448	2.831	0.107
petroMagDiff_i_z	640 432	0.000	620 422	-4.427	0.190	0.244	0.270	0.326	3.686	0.151



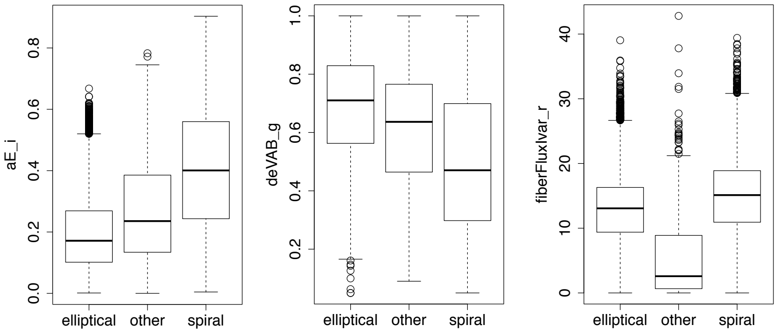
(a) FIBER2FLUXIVAR_U (b) SKYIVAR_R (c) LNLSTAR_R

Figure: Histograms of a selection of features from the SDSS dataset.



(a) EXPRAD_R

Figure: Histograms of the EXPRAD_R feature by target feature level.



(a) AE_I (b) DEVAB_G (c) FIBERFLUXIVAR_R

Figure: Small multiple box plots (split by the target feature) of some of the features from the SDSS ABT.

Modeling

k nearest neighbor model (classification accuracy: 82.912%, average class accuracy: 54.663%)

		Prediction			
		'elliptical'	***'spiral'***	***'other'***	**Recall**
	'elliptical'	115 438	10 238	54	91.814%
Target	***'spiral'***	19 831	50 368	18	71.731%
	'other'	2 905	1 130	18	0.442%

logistic regression model (classification accuracy: 86.041%, average class accuracy: 62.137%)

		Prediction			
		'elliptical'	*'spiral'*	*'other'*	**Recall**
	'elliptical'	115 169	10 310	251	91.600%
Target	***'spiral'***	13 645	56 321	251	80.209%
	'other'	2 098	1 363	592	14.602%

support vector machine model (classification accuracy: 85.942%, average class accuracy: 58.107%)

		Prediction			
		'elliptical'	*'spiral'*	*'other'*	**Recall**
Target	*'elliptical'*	114 721	10 992	18	91.244%
	'spiral'	13 089	57 092	36	81.307%
	'other'	2 654	1 327	72	1.770%

k nearest neighbor model (classification accuracy: 73.965%)

		Prediction			
		'elliptical'	***'spiral'***	***'other'***	**Recall**
	'elliptical'	23 598	4 629	5 253	70.483%
Target	***'spiral'***	4 955	24 734	3 422	74.700%
	'other'	3 209	4 572	25 628	76.711%

logistic regression model (classification accuracy: 78.805%)

		Prediction			
		'elliptical'	***'spiral'***	***'other'***	**Recall**
	'elliptical'	25 571	4 203	3 706	76.378%
Target	***'spiral'***	3 677	26 267	3 166	79.331%
	'other'	2 684	3 763	26 963	80.705%

support vector machine model (classification accuracy: 78.226%)

		Prediction			
		'elliptical'	*'spiral'*	*'other'*	**Recall**
	'elliptical'	24 634	4 756	4 089	73.579%
Target	***'spiral'***	3 763	26 310	3 038	79.460%
	'other'	2 584	3 550	27 275	81.640%

k nearest neighbor model (classification accuracy: 85.557%, average class accuracy: 57.617%)

		Prediction			
		'elliptical'	***'spiral'***	***'other'***	**Recall**
	'elliptical'	116 640	9 037	54	92.770%
Target	***'spiral'***	15 833	54 366	18	77.426%
	'other'	2 815	1 130	108	2.655%

logistic regression model (classification accuracy: 88.829%, average class accuracy: 67.665%)

		Prediction			
		'elliptical'	*'spiral'*	*'other'*	**Recall**
	'elliptical'	117 339	8 302	90	93.326%
Target	***'spiral'***	10 812	59 297	108	84.448%
	'other'	1 757	1 273	1 022	25.221%

support vector machine model (classification accuracy: 87.188%, average class accuracy: 60.868%)

		Prediction			
		'elliptical'	*'spiral'*	*'other'*	**Recall**
	'elliptical'	115 152	10 561	18	91.586%
Target	***'spiral'***	11 243	58 938	36	83.938%
	'other'	2 528	1 237	287	7.080%

The confusion matrix for the 5-level logistic regression model (classification accuracy: 77.528%, average class accuracy: 43.018%).

		Prediction					
		'elliptical'	*'spiral_cw'*	*'spiral_acw'*	*'spiral_eo'*	*'other'*	**Recall**
	'elliptical'	120 625	46	1 515	3 450	95	95.939%
	'spiral_cw'	7 986	373	4 715	2 176	30	2.443%
Target	***'spiral_acw'***	8 395	435	4 928	2 272	35	30.673%
	'spiral_eo'	8 719	75	1 018	28 981	78	74.556%
	'other'	3 038	30	218	619	148	3.660%

The confusion matrix for the logistic regression model that distinguished between only the spiral galaxy types (classification accuracy: 68.225%, average class accuracy: 56.621%).

		Prediction			
		'spiral_cw'	*'spiral_acw'*	*'spiral_eo'*	**Recall**
	'spiral_cw'	5 753	6 214	3 319	37.636%
Target	*'spiral_acw'*	6 011	6 509	3 540	40.528%
	'spiral_eo'	1 143	2 084	35 643	91.698%

The confusion matrix for the 5-level two-stage model (classification accuracy: 79.410%, average class accuracy: 53.118%).

		Prediction					
		'elliptical'	*'spiral_cw'*	*'spiral_acw'*	*'spiral_eo'*	*'other'*	**Recall**
	'elliptical'	117 339	76	2 510	5 716	90	93.326%
	'spiral_cw'	2 354	4 859	5 242	2 802	23	31.799%
Target	***'spiral_acw'***	2 473	5 079	5 499	2 990	25	34.229%
	'spiral_eo'	5 985	965	1 760	30 102	60	77.439%
	'other'	1 757	98	341	834	1 022	25.222%

Evaluation

The confusion matrix for the final logistic regression model on the large hold-out test set (classification accuracy: 87.979%, average class accuracy: 67.305%).

		Prediction			
		'elliptical'	***'spiral'***	***'other'***	**Recall**
	'elliptical'	251 845	19 159	213	92.857%
Target	***'spiral'***	25 748	128 621	262	83.179%
	'other'	4 286	2 648	2 421	25.879%

Deployment

- Jocelyn put the SDSS data through a preprocessing step, standardizing all descriptive features.
- A process was put in place that allowed manual review by SDSS experts to be included in the galaxy classification process — the SDSS processing pipeline flagged any galaxies given low probability predictions for manual review.
- An alert system using the **stability index** was put in place to monitor the performance of the models over time so that any **concept drift** that might take place could be flagged.

1. **Business Understanding**
2. **Data Understanding**
3. **Data Preparation**
4. **Modeling**
   - Baseline Models
   - Feature Selection
   - The 5-level Model
5. **Evaluation**
6. **Deployment**