
FUNDAMENTALS OF MACHINE LEARNING FOR PREDICTIVE DATA ANALYTICS

Algorithms, Worked Examples, and Case Studies

John D. Kelleher
Brian Mac Namee
Aoife D'Arcy

The MIT Press
Cambridge, Massachusetts
London, England



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License

This is an excerpt from the book **Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies** by John D. Kelleher, Brian Mac Namee, and Aoife D'Arcy published by The MIT Press in 2015.

Machine learning is often used to build predictive models by extracting patterns from large datasets. These models are used in predictive data analytics applications including price prediction, risk assessment, predicting customer behavior, and document classification. This introductory textbook offers a detailed and focused treatment of the most important machine learning approaches used in predictive data analytics, covering both theoretical concepts and practical applications. Technical and mathematical material is augmented with explanatory worked examples, and case studies illustrate the application of these models in the broader business context.

After discussing the trajectory from data to insight to decision, the book describes four approaches to machine learning: information-based learning, similarity-based learning, probability-based learning, and error-based learning. Each of these approaches is introduced by a nontechnical explanation of the underlying concept, followed by mathematical models and algorithms illustrated by detailed worked examples. Finally, the book considers techniques for evaluating prediction models and offers two case studies that describe specific data analytics projects through each phase of development, from formulating the business problem to implementation of the analytics solution. The book, informed by the authors many years of teaching machine learning, and working on predictive data analytics projects, is suitable for use by undergraduates in computer science, engineering, mathematics, or statistics; by graduate students in disciplines with applications for predictive data analytics; and as a reference for professionals.

10

Case Study: Galaxy Classification

The history of astronomy is a history of receding horizons.
—Edwin Powell Hubble

Astronomy has gone through a revolution in recent years as the reducing costs of digital imaging has made it possible to collect orders of magnitude more data than ever before. Large-scale sky scanning projects are being used to map the whole of the night sky in intricate detail. This offers huge potential for new science based on this massive data collection effort. This progress comes at a cost, however, as all this data must be labeled, tagged, and cataloged. The old approach of doing all this manually has become obsolete because the volume of data involved is just too large.

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.¹ 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. Although the SDSS has been able to put in place algorithmic solutions to identifying certain objects within the images collected, there have been a number of difficulties. In particular, it has not been possible for the SDSS to develop a solution to automatically categorize galaxies into the different **morphological** groups—for example, spiral galaxies or elliptical galaxies.

This case study² 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. The remainder of this chapter describes the work undertaken by Jocelyn on this project within each phase of the CRISP-DM process.

10.1 Business Understanding

When Jocelyn first arrived at SDSS, she was pleased to find that the business problem she was being asked to help with was already pretty well defined in predictive analytics terms. The SDSS pipeline takes the data captured by the

1 Full details of the SDSS project, which is fascinating, are available at www.sdss.org.

2 Although this case study is based on real data downloaded from the SDSS, the case study itself is entirely fictitious and developed only for the purposes of this book. Very similar work to that described in this section has, however, actually been undertaken, and details of representative examples are given in Section 10.6^[27].

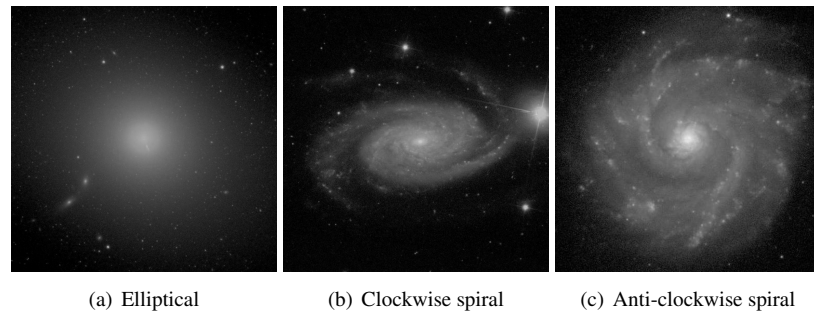
SDSS instruments and processes it, before storing the results of this processing in a centrally accessible database. At the time Jocelyn arrived, the SDSS pipeline included rule-based systems that could classify night sky objects into broad categories—for example, stars and galaxies. SDSS scientists, however, were struggling to build rule-based systems that could accurately perform more fine-grained classifications. In particular, the SDSS scientists wanted a system that could reliably classify galaxies into the important morphological (i.e., shape) types: **elliptical galaxies** and **spiral galaxies**. Classifying galaxies according to **galaxy morphology** is standard practice in astronomy,³ and morphological categories have been shown to be strongly correlated with other important galaxy features. So, grouping galaxies by morphological type is a fundamentally important step in analyzing the characteristics of galaxies.

This was the challenge that the SDSS had hired Jocelyn to address. 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. Although there remained some details left to agree on, the fact that the SDSS had defined their problem in terms of analytics meant that Jocelyn very easily completed the important step of converting a business problem into an analytics solution. Edwin was assigned to Jocelyn as her key scientific contact from SDSS and was eager to answer any questions Jocelyn had as he saw real value in the model she was developing.

The first detail that Jocelyn needed to agree on with Edwin was the set of categories into which sky objects should be categorized. The scientists at SDSS listed two key galaxy morphologies of interest: *elliptical* and *spiral*. The spiral category further divided into *clockwise spiral* and *anti-clockwise spiral* sub-categories. Figure 10.1^[3] shows illustrations of these different galaxy types. Jocelyn suggested that she would first work on the coarse classification of galaxies into elliptical and spiral categories, and then, depending on how this model performed, look at classifying spirals into the more fine-grained categories. Jocelyn also suggested that a third *other* category be included to take into account the fact that all the sky objects labeled as galaxies in the previous step in the SDSS may not actually be galaxies. Edwin agreed with both of these suggestions.

The second detail that Jocelyn needed to agree on with Edwin was the target accuracy that would be required by the system she would build in order

3 This practice was first systematically applied by Edwin Hubble in 1936 (Hubble, 1936).

**Figure 10.1**

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.

for it to be of use to scientists at SDSS. It is extremely important that analytics professionals manage the expectations of their clients during the business understanding process, and agreeing on expected levels of model performance is one of the easiest ways in which to do this. This avoids disappointment and difficulties at later stages in a project. After lengthy discussion, both Jocelyn and Edwin agreed that in order for the system to be useful, a classification accuracy of approximately 80% would be required. Jocelyn stressed that until she had looked at the data and performed experiments, she could not make any predictions as to what classification accuracy would be possible. She did, however, explain to Edwin that because the categorization of galaxy morphologies is a somewhat subjective task (even human experts don't always fully agree on the category that a night sky object should belong to), it was unlikely that classification accuracies beyond 90% would be achievable.

Finally, Edwin and Jocelyn discussed how fast the model built would need to be to allow its inclusion in the existing SDSS pipeline. Fully processed data from the SDSS pipeline is available to scientists approximately one week after images of night sky objects are captured by the SDSS telescopes.⁴ The system that Jocelyn built would be added to the end of this pipeline because it would require outputs from existing data processing steps. It was important that the

⁴ In an interesting example of the persistence of good solutions using older technology, the data captured by the telescopes at the SDSS site in New Mexico is recorded onto magnetic tapes that are then couriered to the Feynman Computing Center at Fermilab in Illinois, over 1,000 miles away. This is the most effective way to transport the massive volumes of data involved!

model Jocelyn deployed not add a large delay to data becoming available to scientists. Based on the expected volumes of images that would be produced by the SDSS pipeline, Jocelyn and Edwin agreed that the model developed should be capable of performing approximately 1,000 classifications per second on a dedicated server of modest specification.

10.1.1 Situational Fluency

The notion of **situational fluency**⁵ is especially important when dealing with scientific scenarios. It is important that analytics professionals have a basic grasp of the work their scientific partners are undertaking so that they can converse fluently with them. The real skill in developing situational fluency is determining how much knowledge about the application domain the analytics professional requires in order to complete the project successfully. It was not reasonable, nor necessary, to expect that Jocelyn would become fully familiar with the intricacies of the SDSS and the astronomy that it performs. Instead, she needed enough information to understand the key pieces of equipment involved, the important aspects of the night sky objects that she would be classifying, and the key terminology involved.

While complex scientific scenarios can make this process more difficult than is the case for more typical business applications, there is also the advantage that scientific projects typically produce publications clearly explaining their work. These kinds of publications are an invaluable resource for an analytics professional trying to come to grips with a new topic. Jocelyn read a number of publications by the SDSS team⁶ before spending several sessions with Edwin discussing the work that he and his colleagues did. The following short summary of the important things she learned illustrates the level of situational fluency required for this kind of scenario.

The SDSS project captures two distinct kinds of data—images of night-sky objects and **spectrographs** of night sky objects—using two distinct types of instrument, an imaging camera and a spectrograph.

⁵ See Chapter ??^[??].

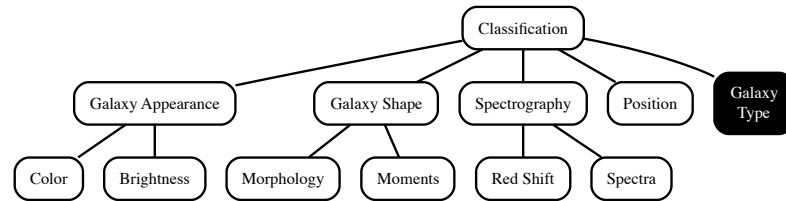
⁶ Stoughton et al. (2002) provides an in-depth discussion of the data collected by the SDSS. A shorter overview is provided at skyserver.sdss3.org/dr9/en/sdss/data/data.asp.

The SDSS imaging camera captures images in *five* distinct **photometric bands**:⁷ ultra-violet (*u*), green (*g*), red (*r*), far-red (*i*), and near infra-red (*z*). The raw imaging data captured from the SDSS telescopes is passed through a processing pipeline that identifies individual night sky objects and extracts a number of properties for each object. For galaxy classification, the most important properties extracted from the images are brightness, color, and shape. The measure of brightness used in the SDSS pipeline is referred to as **magnitude**. **Flux** is another measure that attempts to standardize measures of brightness, taking into account how far away different objects are from the telescope. Measures of flux and magnitude are made in each of the five photometric bands used by the SDSS imaging system. To measure the **color** of night sky objects, the flux measured in different photometric bands is compared. The image-based measures of overall galaxy shape are extracted from the images using **morphological** and **moment** image processing operations. These measures capture how well objects match template shapes—although none is accurate enough to actually perform the galaxy morphology prediction itself.

A **spectrograph** is a device that disperses the light emitted by an object into different wavelengths and measures the intensity of the emission of each wavelength—this set of measures is referred to as a **spectrogram**. The SDSS spectrographs perform this task for manually identified night sky objects and produce spectrograms across wavelengths from visible blue light to near-infrared light. Spectrography data may be useful in galaxy classification because different galaxy types are likely to emit different amounts of different light wavelengths, so spectrograms might be a good indicator for galaxy type. Spectrography also allows measurement of **redshift**, which is used to determine the distance of night sky objects from the viewer.

Once Jocelyn felt that she was suitably fluent with the SDSS situation, she proceeded to the Data Understanding phase of the CRISP-DM process so as to better understand the data available.

⁷ Most consumer digital cameras capture full color images by capturing separate images on red, green, and blue imaging sensors and combining these. The colors red, green, and blue are known as **photometric bands**. The photometric bands captured by the SDSS imaging camera are the same as these bands; they are just defined on different parts of the spectrum.

**Figure 10.2**

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

10.2 Data Understanding

Jocelyn's first step in fully understanding the data available to her was to define the **prediction subject**. In this case the task was to categorize galaxies according to morphology, and therefore galaxy made sense as the prediction subject. The structure of the dataset required for this task would contain one row per galaxy, and each row would include a set of descriptive features describing the characteristics of that galaxy object and a target feature indicating the morphological category of the galaxy object.

Based on her understanding of the SDSS process, Jocelyn sketched out the first draft of the domain concepts diagram for the galaxy classification problem shown in Figure 10.2^[6]. Jocelyn felt that the important **domain concepts** were likely to be the target (galaxy type), galaxy appearance measures (e.g., color), spectrography information (e.g., red shift), and position information (the position of each object in the night sky was also available from the SDSS pipeline). Data with which to implement features based on these domain concepts would likely come from the raw camera imaging and spectrograph images themselves, or from the results of the SDSS processing pipeline.

Jocelyn took this first domain concept draft along to a meeting with Ted, the SDSS chief data architect, to discuss the data resources that would be available for model building. Ted quickly made two observations. First, the spectrograph data collected by the SDSS telescopes was not nearly as extensive as the camera imaging data collected—while there was imaging data for millions of galaxies, there were spectrograms for only hundreds of thousands. Collecting spectrographic information involves a much more complicated process than capturing imaging data, so it is done for a much smaller portion of

the sky. This was likely to continue to be the case, so any solution that relied on spectrographic data as well as imaging data to classify galaxy types would work for only a fraction of the observations made by the SDSS telescopes.

Ted's second observation was that, although there was a huge amount of data available on past observations of night sky objects, only a tiny fraction of these contained manual labels indicating the morphological category to which they belonged. This meant that the data available at the SDSS did not contain a suitable target feature that Jocelyn could use to train prediction models. This is a very common scenario and a real thorn in the side of the predictive model builder—although there is often an almost endless amount of data available for training, little or none of it is labeled with the relevant target feature, making it effectively useless.

Jocelyn's options at this stage were (1) to embark on a large-scale manual data labeling project for which she would hire experts to manually label a suitably large set of historical night sky object observations, or (2) to find some other data source that she could add to the SDSS data to use as a target feature. While the first option is often used, Jocelyn was lucky that another data source became available. Through conversations with Edwin, Jocelyn became aware of a parallel project to the SDSS that offered an intriguing solution to her problem. **Galaxy Zoo**⁸ is a **crowdsourced, citizen science** effort in which people can log onto a website and categorize images of galaxies—taken from the SDSS—into different groups. The Galaxy Zoo project started in 2007 and since then has collected millions of classifications of hundreds of thousands of galaxies.

The galaxy types that Galaxy Zoo citizen scientists could choose from were *elliptical*, *clockwise spiral*, *anti-clockwise spiral*, *edge-on disk*, *merger*, and *don't know*. The first three types are self-explanatory and match directly with the categories of interest to the SDSS project. An *edge-on disk* is a spiral galaxy viewed from the edge, which makes the direction of the spiral arms unclear. A *merger* is a sky object in which multiple galaxies appear grouped together. Examples were labeled as *don't know* when a Galaxy Zoo participant could not place the object in question into one of the other categories.

⁸ Full details of the Galaxy Zoo project and the data released by it are described in Lintott et al. (2011, 2008). The Galaxy Zoo (www.galaxyzoo.org) project referred to in this example is Galaxy Zoo I.

Table 10.1

The structure of the SDSS and Galaxy Zoo combined dataset.

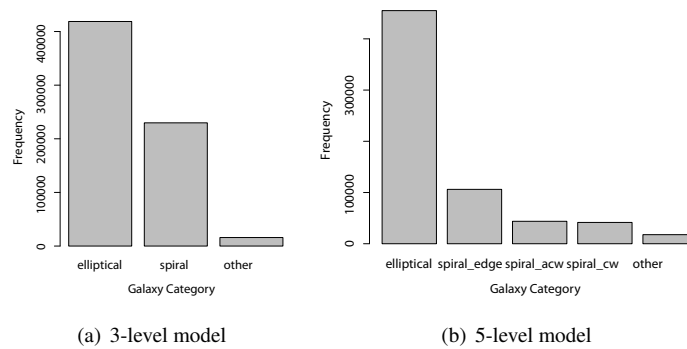
Name	Type	Description
OBJID	Continuous	Unique SDSS object identifier
P_EL	Continuous	Fraction of votes for elliptical galaxy category
P_CW	Continuous	Fraction of votes for clockwise spiral galaxy category
P_ACW	Continuous	Fraction of votes for anti-clockwise spiral galaxy category
P_EDGE	Continuous	Fraction 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

The data from the Galaxy Zoo project was publicly available and therefore easily accessible to Jocelyn. Galaxy Zoo labels were available for approximately 600,000 SDSS galaxies, which Jocelyn felt would be more than enough to use to train and test a galaxy morphology classification model. Conveniently, this also determined the subset of the SDSS dataset (those galaxies used in the Galaxy Zoo project) that Jocelyn would use for this project. With the knowledge that the Galaxy Zoo labels would provide her with a target feature, Jocelyn returned to speak with Ted again about getting access to the SDSS data.

Accessing the results of the SDSS processing pipeline turned out to be reasonably straightforward as it was already collected into a single large table in the SDSS data repository. Ted organized a full download of the SDSS photo imaging data repository for all the objects for which Galaxy Zoo labels existed. This dataset contained 600,000 rows and 547 columns,⁹ with one row for each galaxy observation, containing identifiers, position information, and measures describing the characteristics of the galaxy.

Jocelyn decided to begin her data exploration work by focusing on the target feature. The structure of the data available from the Galaxy Zoo project is shown in Table 10.1^[8]. The category of each galaxy is voted on by multiple Galaxy Zoo participants, and the data includes the fraction of these votes for each of the categories.

9 The fact that the SDSS and Galaxy Zoo make all their data available for free online is a massive contribution to global science. The data used in this case study can be accessed by performing a simple SQL query at skyserver.sdss3.org/dr9/en/tools/search/sql.asp. The query to select all the camera imaging data from the SDSS data release for each of the objects covered by the Galaxy Zoo project along with the Galaxy Zoo classifications is `SELECT * FROM PhotoObj AS p JOIN ZooSpec AS zs ON zs.objid = p.objid ORDER BY p.objid`. Full details of all the data tables available from the SDSS are available at skyserver.sdss3.org/dr9/en/help/docs/tabledesc.asp.

**Figure 10.3**

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

The raw data did not contain a single column that could be used as a target feature, so Jocelyn had to design one from the data sources that were present. She generated two possible target features from the data provided. In both cases, the target feature level was set to the galaxy category that received the majority of the votes. In the first target feature, just three levels were used: *elliptical* (P_EL majority), *spiral* (P_CW, P_ACW, or P_EDGE majority), and *other* (P_MG or P_DK majority). The second target feature allowed three levels for spiral galaxies: *spiral_cw* (P_CW majority), *spiral_acw* (P_ACW majority), and *spiral_edge* (P_EDGE majority). Figure 10.3^[9] shows bar plots of the frequencies of the 3-level and the 5-level target features. The main observation that Jocelyn made from these is that galaxies in the dataset were not evenly distributed across the different morphology types. Instead, the *elliptical* level was much more heavily represented than the others in both cases. Using the 3-level target feature as her initial focus, Jocelyn began to look at the different descriptive features in the data downloaded from the SDSS repository that might be useful in building a model to predict galaxy morphology.

The SDSS download that Jocelyn had access to was a big dataset—over 600,000 rows. Although modern predictive analytics and machine learning tools can handle data of this size, a large dataset can be cumbersome when performing data exploration operations—calculating summary statistics, generating visualizations, and performing correlation tests can just take too long.

Table 10.2

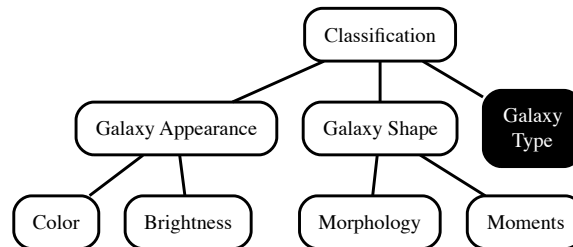
Analysis of a subset of the features in the SDSS dataset.

Feature	Count	%		Min.	1 st		Median	3 rd		Std. Dev.
		Miss.	Card.		Qrt.	Mean		Qrt.	Max.	
RUN	10,000	0.00	380	109.00	2,821.00	3,703.45	3,841.00	4,646.00	8,095.00	1,378.82
RA.1	10,000	0.00	9,964	0.03	151.38	185.26	185.02	220.56	359.99	59.12
DEC.1	10,000	0.00	9,928	-11.23	9.71	24.87	23.41	39.11	69.83	18.92
ROWC_U	10,000	0.00	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ROWC_G	10,000	0.00	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ROWC_R	10,000	0.00	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ROWC_I	10,000	0.00	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ROWC_Z	10,000	0.00	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SKYIVAR_U	10,000	0.00	9,986	-9,999.00	459.81	78.89	798.27	1,083.65	2,197.09	450.26
SKYIVAR_G	10,000	0.00	9,989	-9,999.00	439.55	965.88	2,957.92	6,005.71	9,913.59	2,766.70
SKYIVAR_R	10,000	0.00	9,988	-9,999.00	123.31	201.91	1,091.78	3,347.77	4,623.07	1,514.50
SKYIVAR_I	10,000	0.00	9,986	-9,999.00	46.02	174.79	434.48	1,825.93	2,527.57	851.42
SKYIVAR_Z	10,000	0.00	9,986	-9,999.00	13.60	-234.23	49.57	75.39	205.07	44.51
PSFMAG_U	10,000	0.00	9,768	7.47	20.60	21.08	21.13	21.598	26.19	0.85
PSFMAG_G	10,000	0.00	9,743	8.30	19.06	19.48	19.54	19.967	26.17	0.78
PSFMAG_R	10,000	0.00	9,744	7.45	18.23	18.65	18.68	19.113	26.49	0.76
PSFMAG_I	10,000	0.00	9,744	7.33	17.83	18.27	18.26	18.722	25.46	0.80
PSFMAG_Z	10,000	0.00	9,747	7.40	17.47	17.93	17.90	18.381	23.92	0.82
DEVFLUX_U	10,000	0.00	9,990	-3.68	11.64	43.05	23.07	44.31	28,616.04	194.73
DEVFLUX_G	10,000	0.00	9,987	-1,278.28	48.79	143.71	77.06	133.46	614,662.80	2,401.59
DEVFLUX_R	10,000	0.00	9,983	-4.37	111.04	267.74	152.75	250.65	137,413.00	993.65
DEVFLUX_I	10,000	0.00	9,980	-4.06	160.42	390.98	216.57	351.21	608,862.80	3,041.20
DEVFLUX_Z	10,000	0.00	9,983	-14.72	204.72	528.69	276.99	447.45	2,264,700.00	9,073.95

For this reason, Jocelyn extracted a small sample of 10,000 rows from the full dataset for exploratory analysis using **stratified sampling**.

Given that (1) the SDSS data that Jocelyn downloaded was already in a single table; (2) the data was already at the right prediction subject level (one row per galaxy); and (3) many of the columns in this dataset would most likely be used directly as features in the ABT that she was building, Jocelyn decided to produce a **data quality report** on this dataset. Table 10.2^[10] shows an extract from this data quality report. At this point Jocelyn was primarily interested in understanding the amount of data available, any issues that might arise from missing values, and the types of each column in the dataset.

Jocelyn was surprised that none of the columns had any missing values. Although this is not unheard of (particularly in cases like the SDSS project

**Figure 10.4**

The revised domain concepts diagram for the galaxy classification task.

in which data is generated in a fully automated process) it is very unusual. The minimum values of $-9,999$ for the SKYIVAR_U/G/R/I/Z columns (and some others not shown in Table 10.2^[10]), which were so different from the means for those columns, suggested that maybe there were missing values after all.¹⁰ There were also a number of columns, such as ROWC_U/G/R/I/Z, that had cardinality of 1 (and standard deviations of zero) indicating that every row had the same. These features contained no actual information, so should be removed from the dataset.

Having performed this initial analysis, Jocelyn met again with Edwin and Ted to discuss the data quality issues and, more generally, to review the domain concepts outlined in Figure 10.2^[6] so as to begin designing the actual descriptive features that would populate the ABT. Edwin was broadly in agreement with the set of domain concepts that Jocelyn had developed and was very positive about the use of Galaxy Zoo classifications as a source for generating the target feature. He did explain, however, that Jocelyn's suggestion of using position information was very unlikely to be useful, so that was removed from the set of domain concepts. Edwin also agreed that Ted was correct about the unavailability of spectrograph data for most objects, so this was also removed. The final domain concept diagram is shown in Figure 10.4^[11]. Edwin helped Jocelyn align the columns in the raw SDSS dataset with the different domain concepts, which generated a good set of descriptive features within each domain concept.

Both Edwin and Ted were surprised to see missing values in the data as it was produced through a fully automated process. Simply through eye-balling

¹⁰ Many systems use values like $-9,999$ to indicate that values are actually missing.

the data, Jocelyn uncovered the fact that, in almost all cases, when one suspect $-9,999$ value was present in a row in the dataset, that row contained a number of suspect $-9,999$ values (this was the case for 2% of the rows in the dataset). Although neither Edwin nor Ted could understand exactly how this had happened, they agreed that something had obviously gone wrong in the processing pipeline in those cases and that the $-9,999$ values must refer to missing values.¹¹ **Complete case analysis** was used to entirely remove any rows containing two or more $-9,999$, or missing, values. Before performing this operation, however, Jocelyn first checked that the percentage of missing values was approximately 2% in each of the 3 levels (and in each of the levels in the 5-level model) to ensure that there was no relationship between missing values and galaxy type. There was no obvious relationship, so Jocelyn was confident that removing rows containing missing values would not affect one target level more than the others.

One of the advantages of working in scientific scenarios is that there is a body of literature that discusses how other scientists have addressed similar problems. Working with Edwin, Jocelyn reviewed the relevant literature and discovered a number of very informative articles discussing descriptive features that were likely to be useful in classifying galaxy morphologies.¹² In particular, a number of interesting features that could be derived from the flux and magnitude measurements already in the SDSS dataset were described in the literature. Jocelyn implemented these derived features for inclusion in the final ABT.

In many instances the SDSS dataset contained the same measurement for a night sky object measured separately for each of the five photometric bands covered by the SDSS telescope. Because of this, Jocelyn suspected that there would be a large amount of redundancy in the data as the measurements in the different bands were likely to be highly correlated. To investigate this idea, she generated SPLOM charts for different photometric band versions of a selection of columns from the dataset (see Figure 10.5^[13]), and these showed significant relationships, which confirmed her suspicion. Jocelyn showed these charts to

11 The co-occurrence of multiple missing values in a row is something that it is hard to find through summary analysis of the data and one of the reasons analytics practitioners should always eye-ball extracts from a dataset during the data exploration process.

12 Interested readers might find Tempel et al. (2011), Ball et al. (2004) and Banerji et al. (2010) good references on this topic.

Edwin. Edwin agreed that it was likely that correlations existed between measurements in the different photometric bands but stressed, however, that differences across these bands would exist and might be quite important in predicting galaxy morphology. The existence of a high level of correlation between measurements indicated to Jocelyn that feature selection would be important later during the modeling phase as it had the potential to massively reduce the dimensionality of the dataset.

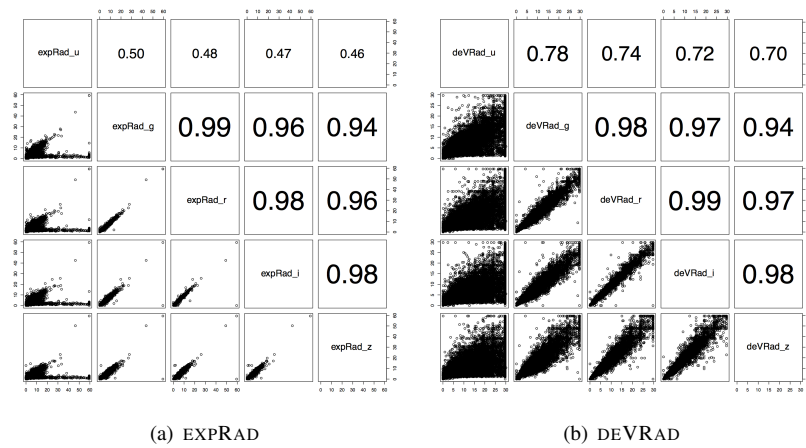


Figure 10.5

SPLOM diagrams of (a) the EXPRAD and (b) DEVRAD measurements from the raw SDSS dataset. Each SPLOM shows the measure across the five different photometric bands captured by the SDSS telescope (*u*, *g*, *r*, *i*, and *z*).

At this point the design of the ABT had fallen into place. For the most part, Jocelyn would use descriptive features directly from the raw SDSS data. These would be augmented with a small number of derived features that the literature review undertaken with Edwin had identified. Jocelyn was now ready to move into the **Data Preparation** phase, during which she would populate the ABT, analyze its contents in detail, and perform any transformations that were required to handle data quality issues.

10.3 Data Preparation

After removing a large number of the columns from the raw SDSS dataset, introducing a number of derived features, and generating two target features,

Table 10.3

Features from the ABT for the SDSS galaxy classification problem.

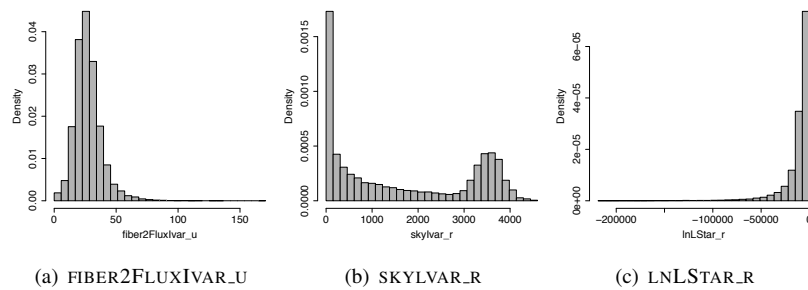
Feature	Feature	Feature
SKYIVAR_U/G/R/I/Z	UERR_U/G/R/I/Z	EXPFLUX_U/G/R/I/Z
PSFMAG_U/G/R/I/Z	ME1_U/G/R/I/Z	EXPFLUXIVAR_U/G/R/I/Z
PSFMAGERR_U/G/R/I/Z	ME2_U/G/R/I/Z	MODELFLUXIVAR_U/G/R/I/Z
FIBERMAG_U/G/R/I/Z	ME1E1ERR_U/G/R/I/Z	CMODELFLUX_U/G/R/I/Z
FIBERMAGERR_U/G/R/I/Z	ME1E2ERR_U/G/R/I/Z	CMODELFLUXIVAR_U/G/R/I/Z
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	LNLSSTAR_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	DEV RAD_U/G/R/I/Z	LNLEDEV_U/G/R/I/Z
PSFFLUXIVAR_U/G/R/I/Z	DEV RADERR_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	PETRO RATIO_I
PETRO RAD_U/G/R/I/Z	EXPRAD_U/G/R/I/Z	PETRO RATIO_R
PETRO RADERR_U/G/R/I/Z	EXPRADERR_U/G/R/I/Z	AE_I
PETRO R50_U/G/R/I/Z	EXPAB_U/G/R/I/Z	PETROMAGDIFF_U_G
PETRO R50ERR_U/G/R/I/Z	EXPABERR_U/G/R/I/Z	PETROMAGDIFF_G_R
PETRO R90_U/G/R/I/Z	EXP MAG_U/G/R/I/Z	PETROMAGDIFF_R_I
PETRO R90ERR_U/G/R/I/Z	EXP MAGERR_U/G/R/I/Z	PETROMAGDIFF_I_Z
Q_U/G/R/I/Z	CMODEL MAG_U/G/R/I/Z	GALAXY_CLASS_3
QERR_U/G/R/I/Z	CMODEL MAGERR_U/G/R/I/Z	GALAXY_CLASS_5
U_U/G/R/I/Z		

Jocelyn generated an ABT containing 327 descriptive features and two target features. Table 10.3^[14] lists these features (features that occur over all five photometric bands are listed as NAME_U/G/R/I/Z to save space).¹³

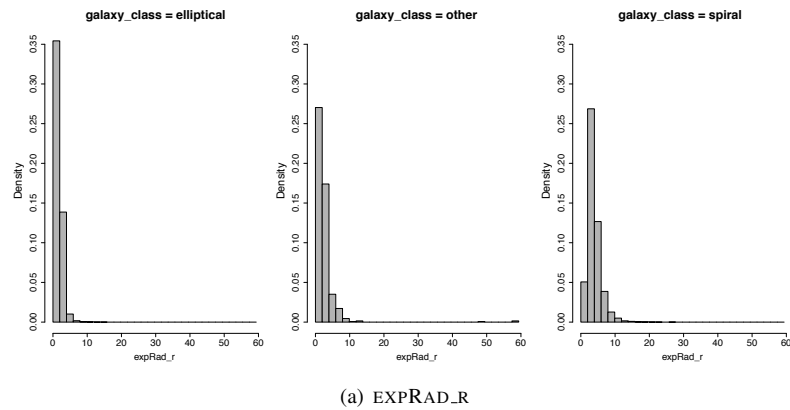
Once Jocelyn had populated the ABT, she generated a data quality report (the initial data quality report covered the data in the raw SDSS dataset only, so a second one was required that covered the actual ABT) and performed an in-depth analysis of the characteristics of each descriptive feature. An extract from this data quality report is shown in Table 10.4^[16].

13 We direct the interested reader to <http://skyserver.sdss3.org/dr9/en/sdss/data/data.asp> for an overview of what these features represent.

The magnitude of the maximum values for the FIBER2FLUXIVAR_U feature in comparison to the median and 3rd quartile value was unusual and suggested the presence of outliers. The difference between the mean and median values for the SKYIVAR_R feature also suggested the presence of outliers. Similarly, the difference between the mean and median values for the LNLSTAR_R feature suggested that the distribution of this feature was heavily skewed and also suggested the presence of outliers. Figure 10.6^[15] shows histograms for these features. The problems of outliers and skewed distributions is clearly visible in these distributions. A number of other features exhibited a similar pattern.

**Figure 10.6**

Histograms of a selection of features from the SDSS dataset.

**Figure 10.7**

Histograms of the EXPRAD_R feature by target feature level.

Table 10.4

A data quality report for a subset of the features in the SDSS ABT.

Feature	Count	% Miss.	Card.	Min.	1 st Qrt.	Mean	Median	3 rd Qrt.	Max.	Std. Dev.
SKYIVAR_U	640,432	0.00	639,983	0.00	465.53	784.78	793.20	1,079.53	2,190.05	447.36
SKYIVAR_G	640,432	0.00	640,081	0.00	442.55	3,318.72	2,949.62	6,008.31	9,898.47	2,769.84
SKYIVAR_R	640,432	0.00	640,178	0.00	127.18	1,629.86	1,094.93	3,342.65	4,596.46	1,513.38
SKYIVAR_I	640,432	0.00	640,042	0.00	48.28	842.18	436.13	1,825.88	2,515.35	852.73
SKYIVAR_Z	640,432	0.00	640,042	0.00	13.90	52.19	49.76	75.10	205.69	44.19
ME2_G	640,432	0.00	629,246	-0.96	-0.13	0.01	0.01	0.15	0.97	0.28
FIBER2FLUXIVAR_U	640,432	0.00	639,827	0.00	20.31	27.24	25.96	32.40	170.70	11.02
PSFMAG_U	640,432	0.00	632,604	13.76	20.59	21.05	21.12	21.58	25.56	0.81
PETROFLUXIVAR_U	640,432	0.00	627,391	0.00	0.16	0.40	0.31	0.53	6.29	0.36
LNLSSTAR_R	640,432	0.00	639,690	-218,875.30	-12,623.05	-12,009.95	-6,771.37	-4,308.99	0.00	16,193.73
PETROMAG_R	640,432	0.00	628,562	11.72	16.76	17.08	17.29	17.61	22.72	0.75
EXPAB_I	640,432	0.00	623,467	0.05	0.49	0.65	0.67	0.81	1.00	0.20
DEREDDIFF_U_G	640,432	0.00	630,319	-2.47	1.29	1.61	1.67	1.89	6.67	0.40
DEREDDIFF_G_R	640,432	0.00	631,627	-1.06	0.64	0.82	0.84	0.99	4.70	0.27
DEREDDIFF_R_I	640,432	0.00	611,597	-4.46	0.36	0.39	0.40	0.44	2.22	0.10
DEREDDIFF_I_Z	640,432	0.00	615,131	-2.29	0.23	0.28	0.30	0.34	5.33	0.11
PETRO_RAT_I	640,432	0.00	640,432	1.12	2.33	2.67	2.68	3.01	25.52	0.46
PETRO_RAT_R	640,432	0.00	640,432	1.18	2.29	2.63	2.64	2.96	10.05	0.42
AE_I	640,432	0.00	640,432	0.00	0.13	0.27	0.23	0.38	0.90	0.18
MODEL_MAGDIFF_U_G	640,432	0.00	630,476	-2.45	1.33	1.65	1.71	1.94	6.83	0.40
MODEL_MAGDIFF_G_R	640,432	0.00	630,437	-1.05	0.68	0.85	0.87	1.03	4.75	0.27
MODEL_MAGDIFF_R_I	640,432	0.00	613,667	-4.46	0.38	0.41	0.42	0.47	2.25	0.10
MODEL_MAGDIFF_I_Z	640,432	0.00	615,346	-2.27	0.25	0.29	0.32	0.35	5.34	0.11
PETROMAGDIFF_G_R	640,432	0.00	631,901	-1.99	0.64	0.83	0.84	1.00	5.13	0.28
PETROMAGDIFF_R_I	640,432	0.00	612,827	-3.32	0.35	0.39	0.41	0.45	2.83	0.11
PETROMAGDIFF_I_Z	640,432	0.00	620,422	-4.43	0.19	0.24	0.27	0.33	3.69	0.15

With Edwin's help, Jocelyn investigated the actual data in the ABT to determine whether the extreme values in the features displaying significant skew or the presence of outliers were due to **valid outliers** or **invalid outliers**. In all cases the extreme values were determined to be valid outliers. Jocelyn decided to use the **clamp transformation** to change the values of these outliers to something closer to the central tendency of the features. Any values beyond the 1st quartile value plus 2.5 times the inter-quartile range were reduced to this value. The standard value of 1.5 times the inter-quartile range was changed to 2.5 to slightly reduce the impact of this operation.

Jocelyn also made the decision to normalize all the descriptive features into **standard scores**. The differences in the ranges of values of the set of descriptive features in the ABT was huge. For example, DEVAB_R had a range as small as [0.05, 1.00] while APERFLUX7IVAR_U had a range as large as [−265,862, 15,274]. Standardizing the descriptive feature in this way was likely to improve the accuracy of the final predictive models. The only drawback to standardization is that the models become less interpretable. Interpretability, however, was not particularly important for the SDSS scenario (the model built would be added to the existing SDSS pipeline and process thousands of galaxy objects per day), so standardization was appropriate.

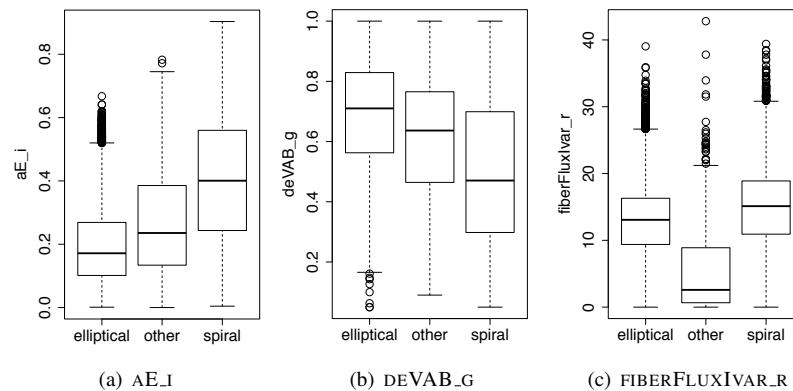


Figure 10.8

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

Jocelyn also performed a simple first-pass feature selection using the 3-level model to see which features might stand out as predictive of galaxy morphology. Jocelyn used the **information gain** measure to rank the predictiveness of the different features in the dataset (for this analysis, missing values were

simply omitted). The columns identified as being most predictive of galaxy morphology were `expRad_g` (0.3908), `expRad_r` (0.3649), `deVRad_g` (0.3607), `expRad_i` (0.3509), `deVRad_r` (0.3467), `expRad_z` (0.3457), and `mRrCc_g` (0.3365). Jocelyn generated histograms for all these features compared to the target feature—for example, Figure 10.7^[15] shows the histograms for the `EXPRAD_R` feature. It was encouraging that in many cases distinct distributions for each galaxy type were apparent in the histograms. Figure 10.8^[17] shows small multiple box plots divided by galaxy type for a selection of features from the ABT. The differences between the three box plots in each plot gives an indication of the likely predictiveness of each feature. The presence of large numbers of outliers can also be seen.

10.4 Modeling

The descriptive features in the SDSS dataset are primarily continuous. For this reason, Jocelyn considered trying a similarity-based model, the ***k* nearest neighbor**, and two error-based models, the **logistic regression** model and the **support vector machine**. Jocelyn began by constructing a simple baseline model using the 3-level target feature.

10.4.1 Baseline Models

Because of the size of the ABT, Jocelyn decided to split the dataset into a **training set** and a large **hold-out test set**. Subsets of the training set would be also used for **validation** during the model building process. The training set consisted of 30% of the data in the ABT (approximately 200,000 instances), and the test set consisted of the remaining 70% (approximately 450,000 instances).¹⁴ Using the training set, Jocelyn performed a 10-fold cross validation experiment on models trained to use the full set of descriptive features to predict the 3-level target. These would act as baseline performance scores that she would try to improve upon. The classification accuracies achieved during the cross validation experiment were 82.912%, 86.041%, and 85.942% by the *k* nearest neighbor, logistic regression, and support vector

¹⁴ It is more common to split an ABT in the opposite proportions (70% for the training set and 30% for the test set). In this case, however, because the ABT was so large it was more useful to have a very large test sample as 200,000 instances should be more than enough for the training set.

Table 10.5

The confusion matrices for the baseline models.

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

		Prediction			Recall
		<i>elliptical</i>	<i>spiral</i>	<i>other</i>	
Target	<i>elliptical</i>	115,438	10,238	54	91.814%
	<i>spiral</i>	19,831	50,368	18	71.731%
	<i>other</i>	2,905	1,130	18	0.442%

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

		Prediction			Recall
		<i>elliptical</i>	<i>spiral</i>	<i>other</i>	
Target	<i>elliptical</i>	115,169	10,310	251	91.600%
	<i>spiral</i>	13,645	56,321	251	80.209%
	<i>other</i>	2,098	1,363	592	14.602%

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

		Prediction			Recall
		<i>elliptical</i>	<i>spiral</i>	<i>other</i>	
Target	<i>elliptical</i>	114,721	10,992	18	91.244%
	<i>spiral</i>	13,089	57,092	36	81.307%
	<i>other</i>	2,654	1,327	72	1.770%

machine models respectively. The confusion matrices from the evaluation of these models are shown in Table 10.5^[19].

These initial baseline results were promising; however, one key issue did emerge. It was clear that the performance of the models trained using the SDSS data was severely affected by the **target level imbalance** in the data—there were many more examples of the *elliptical* target level than either the *spiral* or, especially, the *other* target level. Having a dominant target level, like the *elliptical* target level in this example, means that models trained on this data can over-compensate for the majority target level and ignore the minority ones. For example, based on the confusion matrix in Table 10.5(c)^[19], the misclassification rate for the *elliptical* target level is only 8.756%, while for the *spiral* target level, it is higher, at 18.693%, and for the *other* target level, it is a fairly dire 98.230%. The single classification accuracy performance measure hides this poor performance on the minority target levels. An average class accuracy

performance measure, however, brings this issue to the fore. The average class accuracy scores achieved by the models were 54.663%, 62.137%, and 58.107% by the k nearest neighbor, logistic regression, and support vector machine models respectively. Jocelyn decided to build a second set of models in which she would address the target level imbalance issue.

The target level imbalance in the SDSS dataset arises through **relative rarity**.¹⁵ In the large SDSS dataset, there are plenty of galaxies in the *other* and *spiral* categories; there are just many more in the *elliptical* category. In this case, Jocelyn addressed the target level imbalance problem by using **under-sampling** to generate a new training dataset in which all three target levels had an equal distribution. This was referred to as the **under-sampled training set**. Jocelyn performed the same baseline test on the three model types using this new dataset. The resulting confusion matrices are shown in Table 10.6^[21].

The resulting classification accuracies (average class accuracies and classification accuracies are the same in this case because the dataset is balanced) from the 10-fold cross validation experiment were 73.965%, 78.805%, and 78.226% for the k nearest neighbor, logistic regression, and support vector machine models respectively. The overall performance on this balanced dataset was not as good as the performance on the original dataset; however, balancing the training set did result in the performance on each target level being more equal. Predictions for the *other* target level are actually being performed this time, whereas in the previous example, this target level was essentially being ignored. Choosing between models in this sort of scenario is difficult as it really comes down to balancing the needs of the application—when the system makes errors (as it inevitably will from time to time), what error is least bad? In this example, is it better to classify a galaxy that should be *other* as an *elliptical* galaxy or vice versa? Jocelyn discussed this issue and the results of these two baseline experiments with Edwin, and both decided that it would be best to pursue the optimal performance measured by overall classification accuracy because, in practice, the important thing for the SDSS system was to classify *elliptical* and *spiral* galaxies as accurately as possible.

15 Target level imbalance typically arises through either **absolute rarity** or **relative rarity** of the minority target levels. Absolute rarity refers to scenarios in which there simply do not exist many examples of the minority target levels—for example, in automated inspection tasks on production lines, it is often the case that there are simply very few examples of defective products that can be used for training. Relative rarity, on the other hand, refers to scenarios in which the proportion of examples of the majority target levels in a dataset is much higher than the proportion of examples of the minority target level, but there is actually no shortage of examples of the minority target level.

Table 10.6

The confusion matrices showing the performance of models on the under-sampled training set.

(a) k nearest neighbor model (classification accuracy: 73.965%)

		Prediction			Recall
		<i>elliptical</i>	<i>spiral</i>	<i>other</i>	
Target	<i>elliptical</i>	23,598	4,629	5,253	70.483%
	<i>spiral</i>	4,955	24,734	3,422	74.700%
	<i>other</i>	3,209	4,572	25,628	76.711%

(b) logistic regression model (classification accuracy: 78.805%)

		Prediction			Recall
		<i>elliptical</i>	<i>spiral</i>	<i>other</i>	
Target	<i>elliptical</i>	25,571	4,203	3,706	76.378%
	<i>spiral</i>	3,677	26,267	3,166	79.331%
	<i>other</i>	2,684	3,763	26,963	80.705%

(c) support vector machine model (classification accuracy: 78.226%)

		Prediction			Recall
		<i>elliptical</i>	<i>spiral</i>	<i>other</i>	
Target	<i>elliptical</i>	24,634	4,756	4,089	73.579%
	<i>spiral</i>	3,763	26,310	3,038	79.460%
	<i>other</i>	2,584	3,550	27,275	81.640%

With these baseline performance measures established, Jocelyn turned her attention to feature selection in an effort to improve on these performance scores.

10.4.2 Feature Selection

In the SDSS dataset, many of the features are represented multiple times for each of the five different photometric bands, and this made Jocelyn suspect that many of these features might be redundant and so ripe for removal from the dataset. **Feature selection** approaches that search through subsets of features (known as **wrapper** approaches) are better at removing redundant features than rank and prune approaches because they consider groups of features together. For this reason, Jocelyn chose to use a **step-wise sequential search** for feature selection for each of the three model types. In all cases overall classification accuracy was used as the fitness function that drove the search. After feature selection, the classification accuracy of the models on the test

Table 10.7

The confusion matrices for the models after feature selection.

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

		Prediction			Recall
		<i>elliptical</i>	<i>spiral</i>	<i>other</i>	
Target	<i>elliptical</i>	116,640	9,037	54	92.770%
	<i>spiral</i>	15,833	54,366	18	77.426%
	<i>other</i>	2,815	1,130	108	2.655%

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

		Prediction			Recall
		<i>elliptical</i>	<i>spiral</i>	<i>other</i>	
Target	<i>elliptical</i>	117,339	8,302	90	93.326%
	<i>spiral</i>	10,812	59,297	108	84.448%
	<i>other</i>	1,757	1,273	1,022	25.221%

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

		Prediction			Recall
		<i>elliptical</i>	<i>spiral</i>	<i>other</i>	
Target	<i>elliptical</i>	115,152	10,561	18	91.586%
	<i>spiral</i>	11,243	58,938	36	83.938%
	<i>other</i>	2,528	1,237	287	7.080%

set were 85.557%, 88.829%, and 87.188% for the k nearest neighbor, logistic regression, and support vector machine models respectively. The resulting confusion matrices are shown in Table 10.7^[22]. In all cases performance of the models improved with feature selection. The best performing model is the logistic regression model. For this model, just 31 out of the total 327 features were selected.¹⁶ This was not surprising given the large amount of redundancy within the feature set.

¹⁶ The features selected were AE_I, APERFLUX7IVAR_G, APERFLUX7IVAR_I, APERFLUX7_U, DERED_U, DEVAB_R, DEVRADERR_Z, DEVRAD_U, DEREDDIFF_G_R, EXPRAD_G, EXPRAD_R, FIBER2FLUXIVAR_Z, FIBER2MAGERR_G, FIBERFLUXIVAR_R, FRACDEV_Z, LNLDEV_G, LNLDEV_R, LNLDEV_U, LNLDEV_Z, MCr4_Z, PETROFLUXIVAR_G, PETROFLUXIVAR_I, PETROR50ERR_R, PETROR50_G, PETROR90_G, PETRORATIO_R, PSFFLUXIVAR_I, PSFMAGERR_R, PSFMAG_R, SKYIVAR_U, and SKYIVAR_Z.

Based on these results, Jocelyn determined that the logistic regression model trained using the reduced set of features was the best model to use for galaxy classification. This model gave the best prediction accuracy and offered the potential for very fast classification times, which was attractive for integration into the SDSS pipeline. Logistic regression models also produce confidences along with the predictions, which was attractive to Edwin as it meant that he could build tests into the pipeline that would redirect galaxies with low confidence classifications for manual confirmation of the predictions made by the automated system.

10.4.3 The 5-level Model

To address the finer grained 5-level (*elliptical*, *spiral_cw*, *spiral_acw*, *spiral_eo*, and *other*) classification task, Jocelyn attempted two approaches. First, she used a 5-target-level model to make predictions. Second, she used a **two-stage model**. In this case the logistic regression model used for the 3-level target feature would first be used, and then a model trained to distinguish only between different spiral galaxy types (*clockwise*, *anti-clockwise*, and *edge-on*) would be used to further classify those galaxy objects classified as *spiral* by the first stage.

Based on the performance of the logistic regression model on the 3-level classification problem, Jocelyn trained a logistic regression classifier on the 5-level dataset and evaluated it using a 10-fold cross validation. Following the same approach as in earlier models, Jocelyn performed feature selection using a **step-wise sequential search** to find the best subset of features for this model. Just 11 features from the full set were selected.¹⁷ The resulting classification accuracy on the best performing model that Jocelyn could build was 77.528% (with an average class accuracy of 43.018%). The confusion matrix from this test is shown in Table 10.8^[24]. The overall accuracy of this model is somewhat comparable with the overall accuracy of the 3-level model. The classifier accurately predicts the type of galaxies with the *elliptical* target level and, to a lesser extent, with the *spiral_eo* target level. The ability of the model to distinguish between clockwise (*spiral_cw*) and anti-clockwise (*spiral_acw*) spiral galaxies, however, is extremely poor.

¹⁷ The features selected were SKYIVAR_U, PETROFLUXIVAR_I, PETROR50ERR_G, DEVRAD_G, DEVRADERR_R, DEVRADERR_I, DEVAB_G, EXPFLUX_Z, APERFLUX7_Z, APERFLUX7IVAR_R, and MODEL MAGDIFF_I_Z.

Table 10.8

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

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

Table 10.9

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			Recall
		<i>spiral_cw</i>	<i>spiral_acw</i>	<i>spiral_eo</i>	
Target	<i>spiral_cw</i>	5,753	6,214	3,319	37.636%
	<i>spiral_acw</i>	6,011	6,509	3,540	40.528%
	<i>spiral_eo</i>	1,143	2,084	35,643	91.698%

To test the two-stage classifier, Jocelyn extracted a small ABT containing only spiral galaxies from the original ABT. Using this new ABT, Jocelyn trained a logistic regression model to distinguish between the three spiral galaxy types (*spiral_cw*, *spiral_acw*, and *spiral_eo*). She used step-wise sequential feature selection again, and this time 32 features were chosen.¹⁸ This model was able to achieve a classification accuracy of 68.225% (with an average class accuracy of 56.621%). The resulting confusion matrix is shown in Table 10.9^[24]. Although it is evident from the confusion matrix that the model could distinguish between the edge-on spiral galaxies and the other two types, it could not accurately distinguish between the clockwise and anti-clockwise spiral galaxies.

¹⁸ The features selected were AE_I, APERFLUX7IVAR_R, CMODELFLUXIVAR_U, DEVABERR_G, DEVABERR_Z, DEVAB_G, DEVAB_I, DEVFLUXIVAR_U, DEVMAGERR_U, DEVRAD_G, DEVRAD_U, DEREDDIFF_U_G, EXPABERR_U, EXPAB_G, EXPMAG_Z, EXPRADERR_U, FIBER2FLUXIVAR_R, FIBER2MAG_I, FIBERFLUXIVAR_G, FIBERFLUX_G, FIBERFLUX_R, FIBERFLUX_Z, LNLDEV_R, MCR4_Z, ME1E1ERR_Z, ME1_U, MODEL-MAGDIFF_R_I, PETROMAGDIFF_R_I, PETRO90_R, PSFMAG_U, SKYIVAR_U, and U_R.

Table 10.10

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

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

In spite of the model's difficulty distinguishing between the clockwise and anti-clockwise spiral galaxies, Jocelyn did perform an evaluation of the **two-stage model**. This model first used a 3-level logistic regression model to distinguish between the *elliptical*, *spiral*, and *other* target levels. Any objects classified as belonging to the *spiral* target level were then presented to a model trained to distinguish between the three different spiral types. The two-stage model achieved a classification accuracy of 79.410%. The resulting confusion matrix is shown in Table 10.10^[25].

Although the performance of the two-stage model was better than the performance of the simpler 5-level model, it still did a very poor job of distinguishing between the different spiral galaxy types. Jocelyn discussed this model with Edwin, and they both agreed that the performance was not at the level required by the SDSS scientists for inclusion in the SDSS processing pipeline. It would most likely be possible to create a model that could distinguish between the clockwise and anti-clockwise spiral galaxies, but this would probably require the calculation of new features based on the application of image processing techniques to the raw galaxy images. Based on the time available to the project, Jocelyn did not pursue this avenue and, in consultation with Edwin, decided to continue with just the 3-level model. The best performing model was the 3-level logistic regression model after feature selection (the performance of this model is shown in Table 10.7(b)^[22]). With this model selected as the best performing approach, Jocelyn was ready to perform the final evaluation experiment.

Table 10.11

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			Recall
		<i>elliptical</i>	<i>spiral</i>	<i>other</i>	
Target	<i>elliptical</i>	251,845	19,159	213	92.857%
	<i>spiral</i>	25,748	128,621	262	83.179%
	<i>other</i>	4,286	2,648	2,421	25.879%

10.5 Evaluation

The final evaluation that Jocelyn performed was in two parts. In the first part, she performed a performance test of the final model selected—the 3-level logistic regression model using the selected feature subset—on the large test dataset mentioned at the beginning of Section 10.4^[18]. This dataset had not been used in the training process, so the performance of the model on this dataset should give a fair indication of how well the model would perform when deployed on real, unseen data. The confusion matrix resulting from this test is shown in Table 10.11^[26]. The classification accuracy was 87.979% (with an average class accuracy of 67.305%), which was similar to performance on the training data and well above the target that Jocelyn and Edwin had agreed on at the beginning of the project.

The purpose of the second part of the evaluation was to encourage confidence in the models that Jocelyn had built amongst the SDSS scientists. In this evaluation, Edwin and four of his colleagues independently examined 200 galaxy images randomly selected from the final test set and classified them as belonging to one of the three galaxy types. A single majority classification was calculated from the five manual classifications for each galaxy. Jocelyn extracted two key measurements by comparing these manual classifications to the classifications made by the model she had built. First, Jocelyn calculated an average class accuracy by comparing the predictions made by her model for the same 200 galaxies with the manual classifications made by the SDSS scientists. The average class accuracy was 78.278%, which was similar to the accuracies measured on the overall test set.

Second, Jocelyn calculated an **inter-annotator agreement** statistic for the manual classifications given by the five SDSS scientists. Using the **Cohen's**

kappa¹⁹ measure of **inter-annotator agreement** to measure how closely the manual classifications matched each other, Jocelyn calculated a measure of 0.6. Jocelyn showed that even the SDSS scientists themselves disagreed on the types of certain galaxies. This is not uncommon in this kind of scenario, in which the classifications have a certain amount of fuzziness around their boundaries—e.g., the exact line between an *elliptical* and a *spiral* galaxy can be hard to define—and led to very interesting discussions for the scientists!

Together the strong performance by the model on the large test dataset and the confidence built through the manual annotation exercise meant that Edwin and his colleagues were happy to integrate the 3-level model into the SDSS processing pipeline.

10.6 Deployment

Once Edwin had approved the models that Jocelyn had built, Jocelyn met again with Ted to begin the process of integrating the models into the SDSS processing pipeline. This was a reasonably straightforward process with just a few issues that needed discussion. First, Jocelyn had put the SDSS data through a preprocessing step, standardizing all descriptive features. The standardization parameters (the mean and standard deviation of each feature) needed to be included in the pipeline so that the same preprocessing step could be applied to newly arriving instances before presenting them to the models.

Second, a process was put in place that allowed manual review by SDSS experts to be included in the galaxy classification process. One of the advantages of using a logistic regression model is that along with classifications, it also produces probabilities. Given that there are three target levels, a prediction probability of approximately 0.333 indicates that the prediction made by the model is really quite unsure. A system was put in place in the SDSS processing pipeline to flag for manual review any galaxies given low probability predictions.

Last, a strategy needed to be put in place to monitor the performance of the models over time so that any **concept drift** that might take place could be flagged. Jocelyn agreed with Ted to put in place an alert system using the

¹⁹ The Cohen's kappa statistic was first described in Cohen (1960). Using the Cohen's kappa statistic, a value of 1.0 indicates total agreement, while a value of 0.0 indicates agreement no better than chance. Values around 0.6 are typically understood to indicate an *acceptable* level of agreement, although the exact nature of what is and is not acceptable is very task dependent.

stability index. This would raise an alert whenever the stability index went above 0.25 so that someone could consider retraining the model.

Bibliography

- Ball, N. M., J. Loveday, M. Fukugita, O. Nakamura, S. Okamura, J. Brinkmann, and R. J. Brunner (2004). Galaxy types in the sloan digital sky survey using supervised artificial neural networks. *Monthly Notices of the Royal Astronomical Society* 348(3), 1038–1046.
- Banerji, M., O. Lahav, C. J. Lintott, F. B. Abdalla, K. Schawinski, S. P. Bamford, D. Andreescu, P. Murray, M. J. Raddick, A. Slosar, A. Szalay, D. Thomas, and J. Vandenberg (2010). Galaxy zoo: Reproducing galaxy morphologies via machine learning. *Monthly Notices of the Royal Astronomical Society* 406(1), 342–353.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1), 34–46.
- Hubble, E. (1936). *The Realm of the Nebulae*. Yale University Press.
- Lintott, C., K. Schawinski, S. Bamford, A. Slosar, K. Land, D. Thomas, E. Edmondson, K. Masters, R. C. Nichol, M. J. Raddick, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg (2011, January). Galaxy Zoo 1: data release of morphological classifications for nearly 900 000 galaxies. *Monthly Notices of the Royal Astronomical Society* 410, 166–178.
- Lintott, C. J., K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. J. Raddick, R. C. Nichol, A. Szalay, D. Andreescu, P. Murray, and J. Vandenberg (2008, September). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 389, 1179–1189.
- Stoughton, C., R. H. Lupton, M. Bernardi, M. R. Blanton, S. Burles, F. J. Castander, A. J. Connolly, D. J. Eisenstein, J. A. Frieman, G. S. Hennessy, R. B. Hindsley, Ž. Ivezić, S. Kent, P. Z. Kunszt, B. C. Lee, A. Meiksin, J. A. Munn, H. J. Newberg, R. C. Nichol, T. Nicinski, J. R. Pier, G. T. Richards, M. W. Richmond, D. J. Schlegel, J. A. Smith, M. A. Strauss, M. SubbaRao, A. S. Szalay, A. R. Thakar, D. L. Tucker, D. E. V. Berk, B. Yanny, J. K. Adelman, J. John E. Anderson, S. F. Anderson, J. Annis, N. A. Bahcall, J. A. Bakken, M. Bartelmann, S. Bastian, A. Bauer, E. Berman, H. Böhringer, W. N. Boroski, S. Bracker, C. Briegel, J. W. Briggs, J. Brinkmann, R. Brunner, L. Carey, M. A. Carr, B. Chen, D. Christian, P. L. Colestock, J. H. Crocker, I. Csabai, P. C. Czarapata, J. Dalcanton, A. F. Davidsen, J. E. Davis, W. Dehnen, S. Dodelson, M. Doi, T. Dombeck, M. Donahue, N. Ellman, B. R. Elms, M. L. Evans, L. Eyer, X. Fan, G. R. Federwitz, S. Friedman, M. Fukugita, R. Gal, B. Gillespie, K. Glazebrook, J. Gray, E. K. Grebel, B. Greenawalt, G. Greene, J. E. Gunn, E. de Haas, Z. Haiman, M. Haldeman, P. B. Hall, M. Hamabe, B. Hansen, F. H. Harris, H. Harris, M. Harvanek, S. L. Hawley, J. J. E. Hayes, T. M. Heckman, A. Helmi, A. Henden, C. J. Hogan, D. W. Hogg, D. J. Holmgren, J. Holtzman, C.-H. Huang, C. Hull, S.-I. Ichikawa, T. Ichikawa, D. E. Johnston, G. Kauffmann, R. S. J. Kim, T. Kimball, E. Kinney, M. Klaene, S. J. Kleinman, A. Klypin, G. R. Knapp, J. Korienek, J. Krolik, R. G. Kron, J. Krzesiński, D. Q. Lamb, R. F. Leger, S. Limmongkol, C. Lindenmeyer, D. C. Long, C. Loomis, J. Loveday, B. MacKinnon, E. J. Mannery, P. M. Mantsch, B. Margon, P. McGehee, T. A. McKay, B. McLean, K. Menou, A. Merelli, H. J. Mo, D. G. Monet, O. Nakamura, V. K. Narayanan, T. Nash, J. Eric H. Neilsen, P. R. Newman, A. Nitta, M. Odenkirchen, N. Okada, S. Okamura, J. P. Ostriker, R. Owen, A. G. Pauls, J. Peoples, R. S. Peterson, D. Petravick, A. Pope, R. Pordes, M. Postman, A. Prossapio, T. R. Quinn, R. Rechenmacher, C. H. Rivetta, H.-W. Rix, C. M. Rockosi, R. Rosner, K. Ruthmansdorfer, D. Sandford, D. P. Schneider, R. Scranton, M. Sekiguchi, G. Sergey, R. Sheth, K. Shimasaku, S. Smee, S. A. Snedden, A. Stebbins, C. Stubbs, I. Szapudi, P. Szkody, G. P. Szokoly, S. Tabachnik, Z. Tsvetanov, A. Uomoto, M. S. Vogeley, W. Voges, P. Waddell, R. Walterbos, S. i Wang, M. Watanabe, D. H. Weinberg, R. L. White, S. D. M. White, B. Wilhite, D. Wolfe, N. Yasuda, D. G. York, I. Zehavi, and W. Zheng (2002). Sloan digital sky survey: Early data release. *The Astronomical Journal* 123(1), 485.
- Tempel, E., E. Saar, L. J. Liivamägi, A. Tamm, J. Einasto, M. Einasto, and V. Müller (2011). Galaxy morphology, luminosity, and environment in the sdss dr7. *A&A* 529, A53.