Term Paper: Predictive Model for Agricultural Land Use

Agustin Ferreira, Danilo Mendez, Gian-Piero Lovicu

Introduction

Farmers often have to decide how to deploy land for agricultural production. In particular, they must consider how topography, weather and other environmental factors affect the suitability of their land for particular types of production. Farmers have traditionally made such decisions on the basis of experience, past performance (something worked well, something else didn't) or perhaps on the advice of peers or other professionals. However, the growing availability of highly granular data means that this decision can be more data-driven.

To demonstrate this, we train a neural network to predict agricultural land use. This model can act as a recommender system for farmers. We train the model using a geo-spatial cross-section of agricultural land use (by commodity) in Australia, joined spatially with environmental features related to that particular piece of land. We have chosen Australia to construct our data set because it has a large agricultural sector, varied climate, produces many different types of commodities and has reasonable data availability.

The model predicts agricultural land use by predominant commodity type, as recorded by the Australian Department of Agriculture, Water and Energy (DAWE). In practice the output of the model is a vector of probabilities that represent the suitability of a piece of land for each type of commodity. A crucial assumption in our model is that the classification collected by DAWE, on average, represents the 'best' use for that land. We discuss the credibility of this assumption in the data section. The features mostly relate to climate and topography. The model abstracts from the economic considerations involved in land use decisions (e.g. commodity prices, input availability and prices). Clearly these matter a lot. However, their cyclicality makes it infeasible for them to be included in our model, which focuses on long-term nature of the climatic features. We discuss this more in the data section.

Overall we find that our modelling approach performs well given the data set: accuracy is between 85-90% and we achieve an F1 score of 40-50% (over 74 classes). Our data suffered from class imbalance, so we trained two models: one that corrected for class imbalance by oversampling and one that did not. Although oversampling helped to better predict some minority classes, the improvement was modest and came at the expense of performance in predicting other classes (including the majority classes). For this reason, we decided to retain both models to make recommendations and as an extension, we could ensemble them. We analyse some of the test predictions made by our model and they make intuitive sense given what we know about agriculture, which is reassuring.

Notwithstanding these findings, the reliability of our data requires the reader to interpret our results with some caution. For instance, the climate data were not available over consistent time periods, which is problematic since we expect they are highly correlated over time. The land use data also has some quality concerns. Both of these issues are discussed in detail in the data section.

Theoretical Considerations

When choosing how to best deploy agricultural land it is important to consider the environmental factors that affect production. For a commodity to flourish it has to be well suited to what the region provides, by natural or artificial means. Many factors affect agricultural production and often they are correlated. From an environmental perspective soil, water and temperature are most important (Raja 2021, Pukite 2018). Table 1 lists the features available for each observation in our land use data in more detail.

Selected environmental factors affecting agricultural production:

- **Water access**: this is access to water that is not directly related to rainfall. Water access is determined by proximity to a waterway (which affects soil moisture), the type of that waterway (e.g. river, stream or drain) irrigation access. Water access is also highly correlated with the elevation of a land parcel (i.e. valleys are wetter than ridges).
- **Precipitation:** determines the amount of water directly available for the ecosystem/farm. We include a number of features to capture the overall level of precipitation, its seasonality and its distribution with respect to temperature.
- **Rain days**: the distribution of rain is related to the level of moisture retained in the soil. We measure rain days according to different daily thresholds.
- **Frost days**: some plants are not frost tolerant. Others require a certain number of frost days to produce fruit for the next season (Ramirez and Kallarackal 2015).
- **Sunshine hours**: some plants require sun exposure for maximum productivity (Song and Jin 2020). We include average sunshine hours per day for each month of the year.
- **Temperature:** temperature is important for plant growth and development. Features related to temperature include levels, distribution in relation to precipitation, temperature ranges.
- **Soil type:** the prevailing soil content of an area is important for nutrient acquisition and soil humidity/drainage. Our data include the content levels of clay, silt, sand and organic carbon in the soil.
- **Elevation:** is correlated with temperature, soil moisture and water access.
- **Area:** some commodities may require a larger land area for production to be feasible.

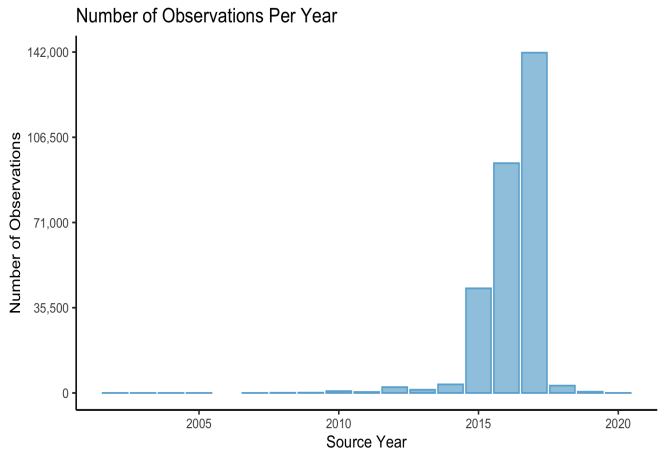
Data

Land use

The Australian Department of Agriculture, Water and Energy (DAWE) is responsible for collecting the land use data (ABARES 2021). The data is updated continuously based on submissions by each of the Australian states and compiled by DAWE into a vector file at the

national level. Data resolution and timeliness therefore depend on state submissions, as well as the remoteness of particular areas.

Land use data for most of Australia were submitted in the past 5 years, except for South Australia and some parts of Western Australia and Queensland. Overall, the data contained in these submissions were collected between 2002 and 2019, though most of the observations are from 2015-2017 (Graph 1). This introduces some temporal randomness in the data collection. Data resolution varies by state, but is positively correlated with remoteness. For example, the Northern Territory contains some of the most remote parts of Australia (Figure 1).



Graph 1: Land Use observations by year of collection

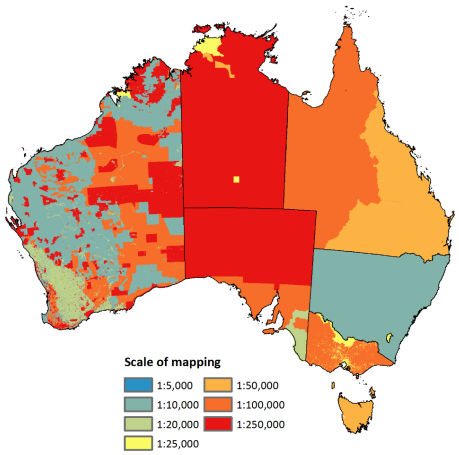
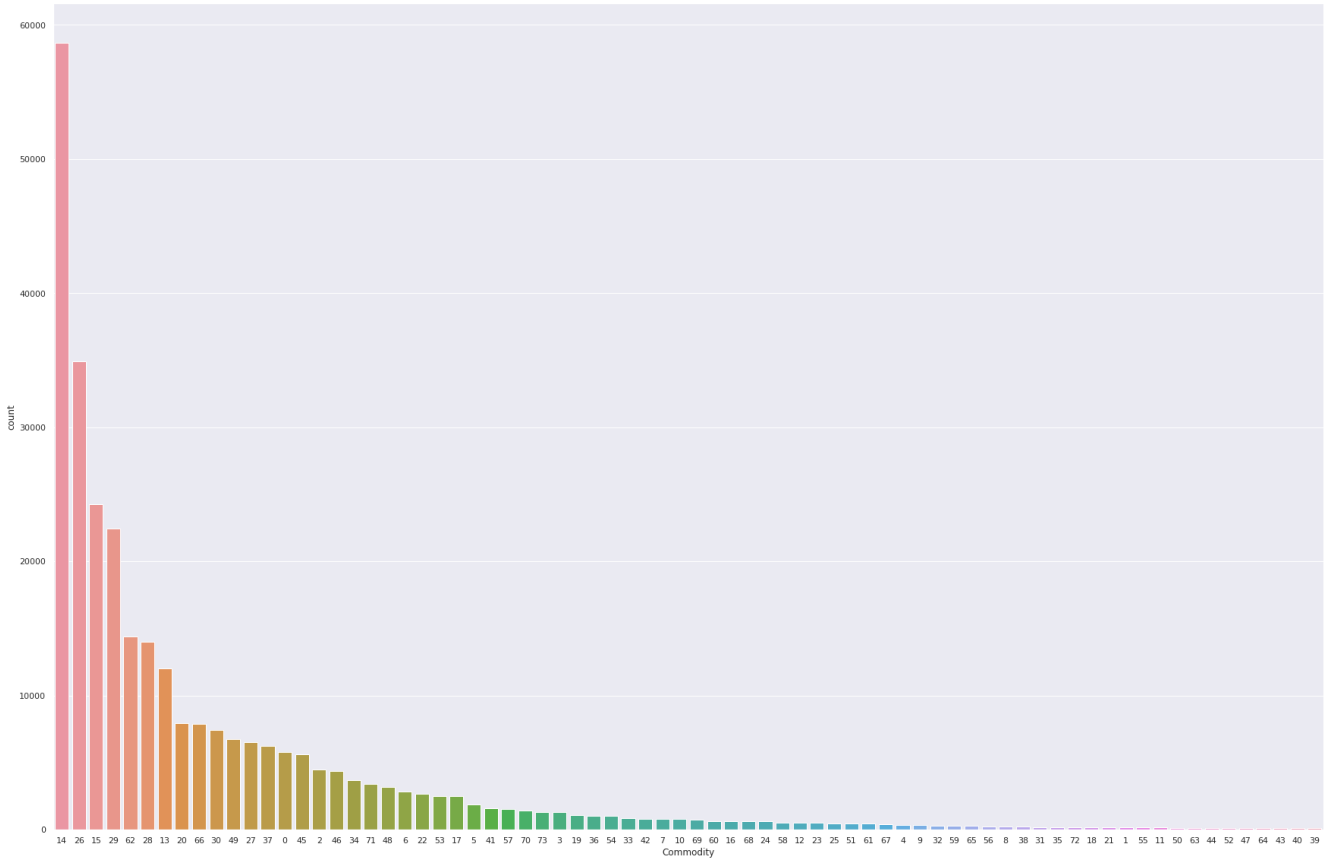


Figure 1: Scale profile of data

Land classified for agricultural use is a subset of the overall land use data and is split into polygons according to landowner and the dominant commodity produced (see Attachment A for an example). DAWE infers commodity type from a combination of satellite data and information collected in the field (e.g. landowner surveys). Of course, it is important to recognise that this classification occurs at a point-of-time (which differs by observation) and that land use has the potential to vary year on year depending on climate, market conditions, etc. If we can assume that on average only a small fraction of agricultural land changes its main use each year, the timing of data collection is unlikely to affect our results too much.

Despite issues with time consistency, this is an impressive dataset. It contains approximately 300,000 unique polygons classified by commodity type. Even though we only observe one type of land use for each observation, with such a large dataset (and the temporal randomness associated with the data collection), we can have some confidence that many commodities will have at least some representation. Then, it is not an issue if a piece of land might be suitable for multiple commodities. So long as we have some observations for each class and the feature set is similar for them, then using a classification model is valid. In other words, we assume that the conditional probability of observing a commodity given a set of features from our data is similar for commodities that are equally suitable for a given piece of land.

After removing classes with a very small number of observations (set at 100, but this can be tuned), the data contains 74 classes for our classification task. Graph 2 shows a frequency histogram of each of the classes. The most commonly observed classes include cattle (dairy), grapes (wine), cattle (meat), sheep, citrus and sugar cane. Clearly, there are some issues with class imbalance, which we address in the modelling section.



Graph 2: Frequency of land use observations by class

Another important assumption we make is that the land use classified by DAWE is the ‘best use’ for that particular piece of land, given the features we observe in our data set. Clearly this assumption is not satisfied, we cannot assume every landowner in the data set has made an optimal production decision. However, if we can assume that sub-optimal land use is random across our observations (independent of a particular class or combination of features) then we can treat it as a source of noise in our data. This assumption is helped by the fact that the data collection includes an element of temporal randomness that is presumably not related to the type of land use or climatic conditions.

Features

We have compiled features from various sources: the Australian Bureau of Meteorology (BoM), WorldClim, OpenStreetMaps and the International Soil Reference and Information Centre

(ISRIC). Table 1 outlines each variable, including its source, time period and transformations made to the raw data.

Most of the data were available as raster data files, which we projected onto a fishnet of the Australian continent (at a resolution of 0.1 degrees or ~11km). We then intersected each fishnet with the land use data. In many cases there was not an exact overlap, since the resolution of the fishnet is not as granular as the land use polygons. In this instance we assigned the nearest fishnet point to each polygon.

Data on waterways were available as a vector file. For this feature, we calculated the distance from each land use polygon to its nearest waterway. Where this distance was less than 0.002 degrees (~200m), we calculated an additional variable that measured the length of that waterway within a 0.002 degree buffer around the land use polygon (this was set to zero for all other polygons). Additionally, we included the type of waterway as a categorical variable and extracted a polygon's access to irrigation from the description of the commodity type (which includes this information).

Attachment A shows some water access features overlaid on top of land use for cattle farming (as an example, so you can see the data). It shows the disparity in access to water depending on the location of cattle farming operations in different parts of Australia, and highlights the varied climate of the Australian continent. This helps the model generalise.

Feature	**Type**	**Features**	**Period**	**Transform**	**Source**
Precipitation (mm)	raster (2.5 deg)	Annual Wettest month, quarter Driest month, quarter Warmest quarter Coldest quarter Seasonality	1970-2000 (average)	fishnet	WorldClim
Temperature (℃ x 100)	raster (2.5 deg)	Mean annual Annual range Max of warmest month Min of coldest month Mean warmest quarter Mean coldest quarter Mean wettest quarter Mean driest quarter Isothermality Mean diurnal range Isothermality Seasonality	1970-2000 (average) 2070	fishnet	WorldClim
Rain (days)	raster	Annual thresholded at 1, 2, 3,	1961-1990	fishnet	BoM

	(0.1 deg)	5, 10 and 25 mm in a day	(average)		
Frost (days)	raster (0.1 deg)	Annual	1981-2005 (average)	fishnet	BoM
Sunshine (hours)	raster (0.1 deg)	Monthly average (each point has >15 years of data)	Between 1900-2003	fishnet	BoM
Waterways	vector	Distance to nearest (km) Length in buffer (km) Nearest type (categorical)	unknown	distance, buffer	OpenStreetMap
Soil type (depth 5cm)	raster	Silt Clay Sand Organic carbon	2020	fishnet	ISRIC
Elevation	raster (0.5 deg)	Elevation (m)	unknown	fishnet	WorldClim
Irrigation access	NA	Binary if in commodity desc	various	From text	DAWE

One issue with this data is that many of the variables are highly correlated, but were in some cases collected over different time periods. We have assuaged this problem by using long-term averages, which are less volatile and produce a degree of temporal overlap between variables. This of course creates a mis-match with our land use data, which is point-in-time, but this is by construction - we want to model land use decisions on the basis of long-term conditions.

The temporal randomness in the land use data, along with use of long-term climate data motivated the exclusion of the economic features from our data. Economic considerations (e.g. commodity prices, input availability) are often cyclical and volatile, compared to the long-term climate data we have collected. They are also only relevant for land use decisions regarding some classes in our data and can often be independent of environmental factors. For instance, a grain farmer may decide between producing wheat and barley based on economic considerations, which from an environmental perspective are equivalent given their land. In contrast, a farmer deciding whether to plant an avocado orchard is more concerned with environmental factors than economic factors, since the production decision is long-term and cannot easily change once the investment is made. The latter is the type of behaviour we are trying to predict.

Modelling Methodology

Our classification model is a multi-layer perceptron with three hidden layers. The first and second hidden layers have 1024 hidden units, the third has 512. We use one-dimensional batch normalisation in each of the hidden layers and for the second and third layers also apply dropout as a form of regularisation to prevent overfitting (drop out probability is set to 0.2). Our activation function is a standard ReLU for each of the hidden layers and for the fully connected final layer, we use a softmax. Figure 2 displays our model architecture.

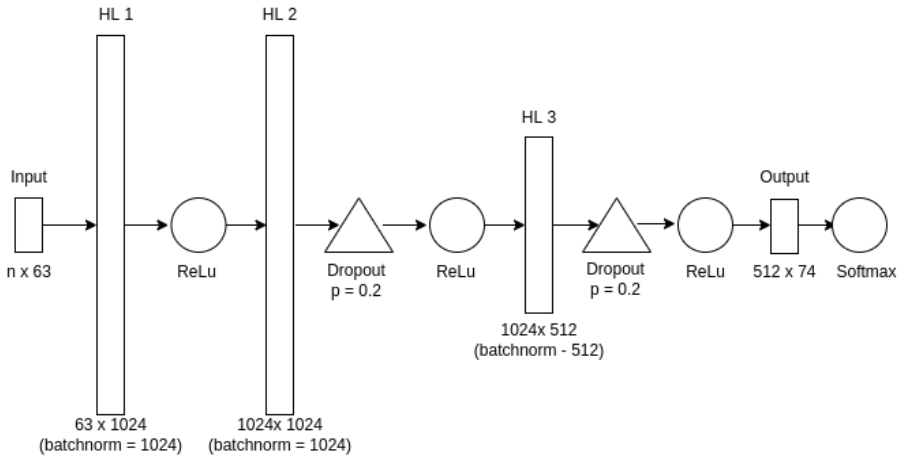


Figure 2: Model architecture

To pre-process our data we use a pipeline to prevent data leakage from our training set to our validation set. For the numerical features, we standardise using a min-max scaler and impute missing values using the median for numerical features and KNN-Imputer for categorical features. For the categorical features, we use a one-hot encoder.

As mentioned earlier, our data set suffers from class imbalance. To address this, we apply an oversampling technique called SMOTE (Synthetic Minority Oversampling Technique) to balance the classes (Chawla 2002). SMOTE looks at the k-nearest neighbours of an under-represented observation and then generates synthetic points along the linear interpolation between them (across all dimensions). The output from SMOTE is a perfectly balanced dataset. The trade-off of applying the oversampling is that our training data is now much larger and consequently the model takes twice as long to train. After applying oversampling, our training data consists of 3.2 million rows.[1] We train two models: one on the original data set and one on the oversampled data set.

To train our models we use a cross-entropy loss function, the Adam optimiser with a learning rate of 0.0007 and a mini-batch size of 1000. We train the models for 30 epochs. We use the classes with the three highest probabilities to calculate model accuracy (so if the model

[1] In some instances, the predicted probabilities from oversampled data are invalid, because they need to be adjusted to reflect the true class weights in the data. This is important when using evaluation metrics that take the probabilities themselves as an input (such as ROC AUC). However, we only use evaluation metrics that require the *rank* of the predicted probabilities to compare with the true label (e.g. accuracy and precision), so this is not of concern for interpreting our results.

predicted the true class within its top three guesses, we classify it as having made a correct prediction). We made this adjustment to account for the fact that multiple commodities may be simultaneously suitable for a given piece of land. It also reflects how a user of the model would present results - by presenting a handful of options, rather than just a single commodity.

Results

To evaluate our models we use several metrics such as accuracy, precision, recall and the F1 score. We modify the calculation of accuracy based on the top 3 predictions. For the other metrics, we are unable to do this efficiently, so these are just on the basis of the highest probability prediction.

The model trained on the original data achieved a (modified) accuracy of 90.5%, while the model trained on the oversampled data achieved an accuracy of 84.5%. These results are intuitive, in the original data model performance is not penalised much for misclassifying under-represented classes because there are few observations of these. However, in the oversampled data the classes are balanced so we expect more mistakes from the classes that were originally under-represented. The drop in accuracy occurs despite the amount of training data available to the model increasing more than 10-fold.

```
> normal
# A tibble: 4 × 3
  metric    Normal Oversampled
  <chr>      <dbl>       <dbl>
1 Accuracy    90          84.5
2 F1          42.9        49.0
3 Precision   39.5        59.7
4 Recall      63.4        40.9
```

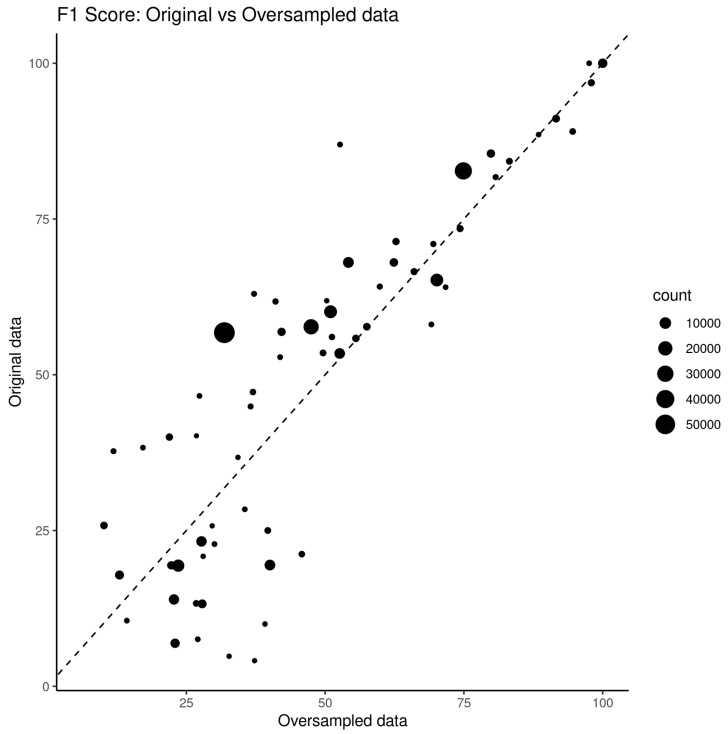
Table 2: Model performance (on test set)

The precision, recall and F1 score are much lower than the accuracy, which makes sense since we are only using the highest probability prediction (we use unweighted metrics, though the weighted versions are not substantially different). There are some interesting differences in the precision and recall between the two models. For the original model the recall is much higher than the precision, while we observed the opposite for the oversampled model. The recall measures the share of true observations the model was able to correctly predict. This is expected, so long as the model predicts the majority classes well, the recall of the original model will be high. However, the precision of predicting each class is much lower since the model is predicting a lot of the minority classes poorly (by predicting them as a majority class the precision of predicting that majority class falls).

For the oversampled model, we observe the opposite - the precision is much higher than the recall. This also makes sense if we flip the logic we just applied. There are no majority classes in the oversampled data, so to achieve a high recall score the model must predict *all* of the classes well. Precision is much higher though, since the model is making fewer mistakes among

the (previously) minority classes. Taking the harmonic mean of these two metrics gives the F1 score, which suggests the oversampled model performs better.

However, we must dig a little deeper into the F1 scores. Graph 3 plots the F1 score for each class in each model. The size of the points measure the number of observations for each class (in the original model). Classes above the 45 degree line are predicted better by the original model, below the line they are predicted better by the oversampled model. It is clear from the graph that the original model performs better at predicting the majority classes and in general does better or equally as well for most classes. However, we can see the oversampled model outperforms for a number of minority classes, though the absolute scores are still very low. This is possibly because we are oversampling from such a small pool of observations that the additional performance we gain from the technique is modest (even with the perturbations introduced by the SMOTE algorithm).



Graph 3: F1 Scores by Class

Overall this tells us that in an absolute sense, there is not conclusive evidence to prefer one of the models over the other. Since our model is a recommender, we decide to take the output from the two models as equally valid. Thus we will assess predictions from both models when providing the recommendations for new plots of agricultural land.

Since the results from the models are complementary, we could create an ensemble of both models and consider predictions from that. In this instance, we did not use an ensemble because we ran out of free GPUs on Google colab.

Analysis

Analysing the model from a technical perspective is important. However, we would also like to check if the predictions from our model are intuitive and consistent with our prior knowledge of agriculture. We expect two things from our predictions: that they are agriculturally compatible and that they align with the ecological reality of the land (our features). Agriculturally compatible means that the predicted commodities require similar conditions to flourish (i.e. we don't predict wheat and sugarcane or almond and mangoes for a given plot). This is especially important for non-irrigated farmland, which relies wholly on the ecological context of the land. For irrigated farmland there is more flexibility (since water access is guaranteed), so we focus more on features like temperature and elevation. As an additional exercise, we also look at how changing the irrigation status of a given observation affects the commodities the model predicts.

We perform this exercise on two random observations selected from our test set. The first observation is in Victoria (in the south east of Australia). This observation has a minimum temperature above freezing, low annual rainfall and is small. Based on these conditions, we expect the model to predict commodities suitable for dry and warm weather conditions.

The original model predicts: sheep (90%), chicken (7%), horses (0.7%) with no access to irrigation. These choices seem very reasonable for the ecological context of the plot and share sufficient characteristics to be agriculturally compatible. The oversampled model predicts: chickens (97%), sheep (2%) and blackberries (0.01%). Again, the first commodities are appropriate for the weather conditions, but blackberries require more water to be profitable (though it has a negligible probability of prediction).

With access to irrigation, the only change is that we no longer expect dry climate commodities (even though they are still appropriate). The original model predicts oil Mallee (eucalyptus oil) (22%), pistachios (18%) and sheep (10%). All three of these commodities are appropriate given the context. Interestingly oil Mallee is produced from the native eucalyptus tree, which is mostly found in Western Australia. For the oversampled model we get: oil Mallee (eucalyptus oil) (39%), cattle dairy (35%) and almonds (12%), which are again all appropriate.

It is important to note that even in the case of the original model, we are still predicting rare classes (i.e. pistachios), which is a very good sign. In addition, there is some overlap between all models, which is a good consistency check. Overall, the diversity of predictions from our models leaves the farmer or extension agent with a good combination of recommendations.

The second example is in South Australia. This farm is in a very dry and warm location with 275 mm of rain a year (on average). We are expecting more small animal agriculture and perennial systems here since these tolerate dryness and heat better than large cattle or annual production. Table 3 summarises the predictions under each of the four models.

Original	Irrigated	Sheep (51%), Chicken Eggs (40%),

		Nectarines (5%)
Original	Non-irrigated	Horses (69%), Blackberries (19%), Melons (5%)
Oversampled	Irrigated	Chicken Eggs (63%), Nectarines (21%), Sheep (14%)
Oversampled	Non-irrigated	Horses (88%), Blackberries (6%), Grapes (4%)

Table 3: Analysing an observation

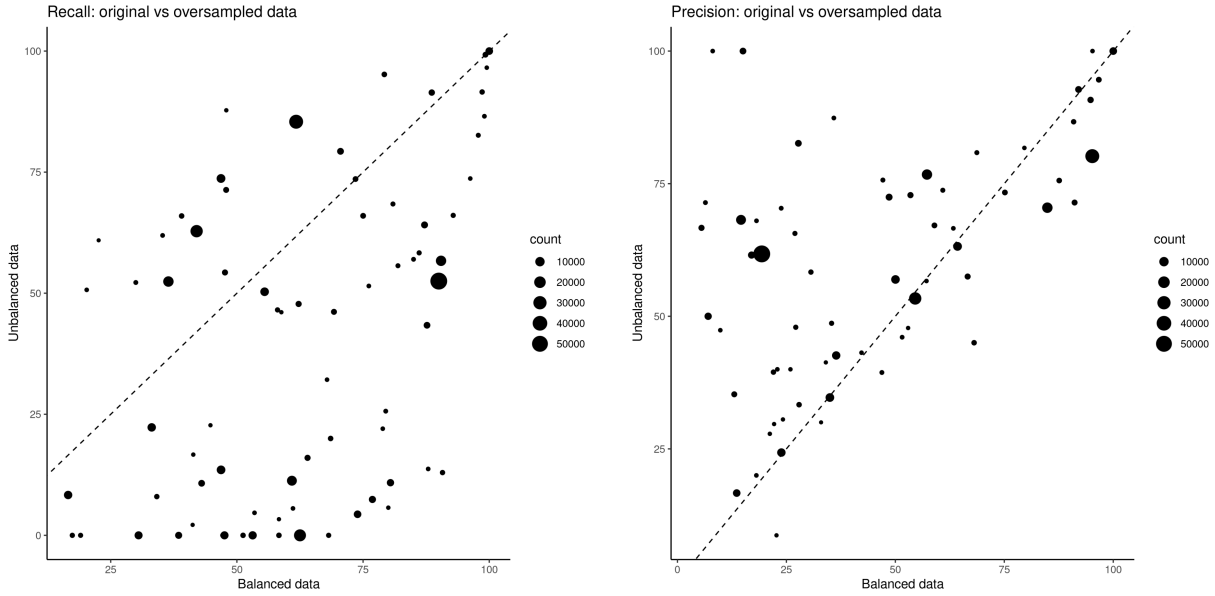
These examples also make some intuitive sense, given our prior. Sheep, chicken (for certain breeds) and horses can do well in arid climates (if enough feed is provided). Also notice that, relative to the previous example, the probabilities are more sensitive to irrigation access than the model being used. Introducing irrigation causes the model to recommend nectarines, which do well in warm climates but also need access to water. The presence of melons and blackberries (although again with low probability relative to the first prediction) demonstrates that the model is not foolproof and would still benefit from expert judgement.||

Conclusion

In this paper, we constructed a data set that spatially joined a database of agricultural land use in Australia to a host of environmental factors. We used this data set to train a multi-layer perceptron to predict agricultural land use. Since our data suffered from class imbalance, we trained two versions of the model: one where we corrected for class imbalance by oversampling and one where we didn't. Although oversampling helped to better predict some minority classes, the improvement was modest and came at the expense of performance in predicting other classes (including the majority classes). For this reason, we decided to retain both models and as an extension, we could ensemble them.

This model can act as a recommender system for farmers considering the commodities it should produce given a piece of land. Although there are some issues with the quality and consistency of our data, the model performs reasonably well and makes intuitive predictions. Extensions to this work would focus on improving the quality of this data set, including obtaining greater representation of minority classes and higher resolution features. One could also consider whether future climate projections could be used to see how climate change may affect the suitability of land for agricultural use.

Appendix



Charts A1 and A2: Precision and Recall between the two models

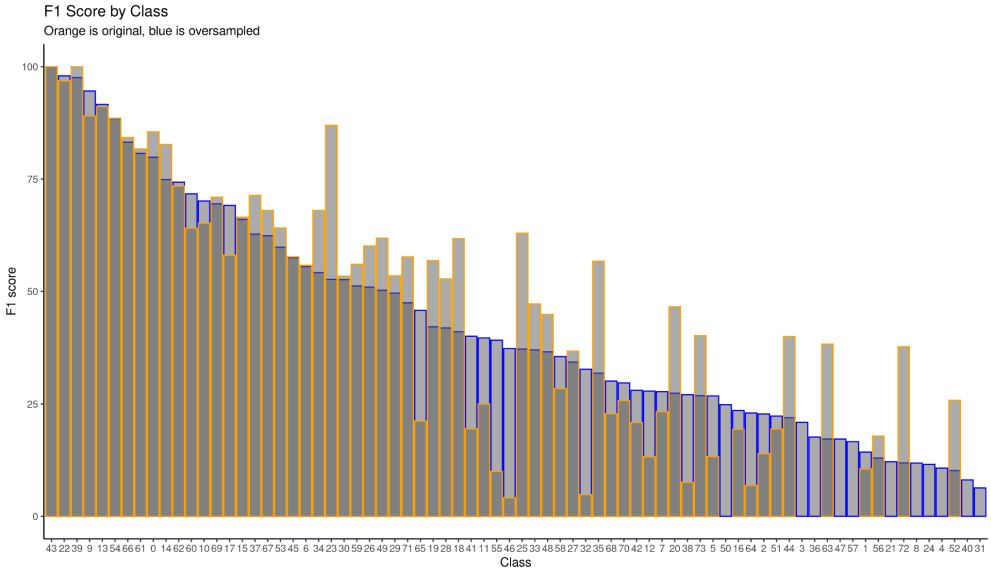


Chart A2: F1 Score by class

References

ABARES 2021, Catchment Scale Land Use of Australia – Commodities – Update December 2020, Australian Bureau of Agricultural and Resource Economics and Sciences, Canberra, February, CC BY 4.0, DOI: 10.25814/jhjb-c072

A. Suruliandi, G. Mariammal & S.P. Raja (2021) Crop prediction based on soil and environmental characteristics using feature selection techniques, Mathematical and Computer Modelling of Dynamical Systems, 27:1, 117-140, DOI: 10.1080/13873954.2021.1882505

V. Cintina, V. Pukite (2018), Analysis of Influencing Factors of Use of Agricultural Land, Rural and Environmental Engineering, Landscape Architecture, DOI: 10.22616/rrd.24.2018.028

Ramírez F., Kallarackal J. (2015) Climate Change and Chilling Requirements. In: Responses of Fruit Trees to Global Climate Change. SpringerBriefs in Plant Science. Springer, Cham. https://doi.org/10.1007/978-3-319-14200-5_9

Song, L.; Jin, J. Effects of Sunshine Hours and Daily Maximum Temperature Declines and Cultivar Replacements on Maize Growth and Yields. *Agronomy* **2020**, *10*, 1862. https://doi.org/10.3390/agronomy10121862

Hausfather, Z. and Peters, G.P., 2020. Emissions–the 'business as usual' story is misleading.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, pp.321-357.