**Predicting corporations’ compliance outcomes in the Peruvian governance market**

**Summary**

**Explanation of project**

I aim to use machine learning to predict the compliance behaviour of listed corporations in Peru when confronted by a new voluntary corporate governance code introduced there in 2014. Corporations that are listed on the *Bolsa de Valores de Lima* (BVL) are required to submit annual returns to the *Superintendencia del Mercado de Valores* (SMV). The reason for this exercise is that there is a wide spread of compliance behaviour, and this analysis is part of a wider study to understand why corporations either do or do not comply with the code.

**Data**

As outcome data, I use aggregated records of corporate compliance submitted to the regulator, the SMV, on an annual basis. These data were obtained from the regulator’s website [www.smv.gov.pe](http://www.smv.gov.pe).

As feature data, I collected data on the 208 listed corporations in 2017 covering 10 aspects of their activities and engagement with financial partner and client markets. This data was collected from a wide range of sources.

Both sets are in accompanying excel.

**Model**

The model I use is a logistic classifier chosen from a list of 6 classifiers offered by scilearn for its stability of results.

**Hyperparametric tuning and feature engineering.**

The data were tuned by reducing the range of features using Lasso regression.

The classifier threshold was set to 0.39 to balance sensitivity and specificity scores.

The model was tuned using penalties and regularization.

**Results**

The model achieved a true positive prediction of 84% true positive for a single rule comprising the code. This methodology can be extended to other rules in due course.

The results are statistically significantly better than random selection

Table of Contents

[I Introduction 2](#_Toc158972507)

[II Methodology 3](#_Toc158972508)

[III Results 9](#_Toc158972509)

[IV Conclusions 17](#_Toc158972510)

# **I Introduction**

**Outline of the context**

To recap on the project, since 2014, corporations in Peru have been required to report on their compliance with a set of rules constituting a code of practice posed as questions relating to aspects of their corporate governance. This initiative has been motivated by the desire of the Peruvian authorities to improve standards of corporate governance both to enhance its dealings with the international business community and also to facilitate its entry into the OECD. The challenge for the authorities is that there is a very wide disparity between high and low compliance corporations and so the picture presented to the outside world is not impressive. However, the authorities are reluctant to exert influence on corporations and indeed are not clear how pressure might be exerted.

I am interested in developing a model to predict the extent to which corporations in Peru comply with the rules set down by a code of conduct for corporate governance with a view to predicting corporation’s compliance behaviour in Peru. Later, I might use this model to predict compliance behaviour in other countries such as Chile, Colombia and Mexico.

**Objectives for machine learning**

My intention is that any model should be used by policy makers to firstly gauge expectations about levels of compliance to be expected from corporations so that they can form a realistic view of which corporations to target to achieve higher overall levels of compliance and so target scarce resources effectively, and second, to make a judgement about the likely impact on compliance of any changes to rules or introduction of new rules into the compliance market.

The likely use of the model will be to identify corporations whose compliance is much lower than expected to enable policy makers and others to seek to influence those corporations to improve their governance. The outputs of the model should therefore be to provide a reasonable expectation given a set of features in order to identify corporations which have consistently lower compliance.

**Dealing with inevitable trade-offs of machine learning**

There are several inevitable trade-offs inherent in machine learning and some of these are directly relevant to the case of a predictive model for the governance market.

The main trade-offs involve that between model performance, whether this be expressed as accuracy or predictive power and the degree to which the model can either be interpreted by practitioners or explained to other stakeholders and its ease of use. As a general guide, more complex models can provide greater accuracy and predictive power but either because of their nature or degree of complexity are more difficult to explain to stakeholders and more difficult for users to operate.

Other trade-offs exist such as the use of ensemble models versus single models; the reliance on automated feature learning versus domain knowledge; and algorithmic complexity versus training time, but these have some commonality with the main axis of accuracy versus explainability.

Given the intended use is for policy-formulation, as a matter of strategy, I will therefore err on the side of explainability, interpretability and ease of use.

# **II Methodology**

**Data classification**

The **dependent variable** data have been obtained from the regulatory body in Peru, the *Superintendencia del Mercado de Valores* which collects and maintains compliance data from all listed corporations in Peru and makes it available on its website.

The data collected consist of responses to 88 questions[[1]](#footnote-1) relating to aspects of corporate governance. Corporations differ significantly in their levels of compliance with the 88 rues comprising the code, though there is a slow trend to improve compliance overall at about one half percent per annum.



Corporations are legally obliged to comply with this data request every year, although they are not required to comply with the behaviours implied by the questions ie they are not required to comply. If they do not comply, they are requested to provide a shore narrative explaining why not. Other considerations include the following:

* The data is available from 2014 to date although the project collected data from 2014 to 2020.
* Once the data is collected and entered onto the SMV’s records, it does not alter
* The form of the data provided is a series of responses to the 88 questions either a “SI” or a “NO”.
* The scale of the available data is given approximately by the following: c 200 corporations reporting on 88 questions for each of 7 years yields 130,000 data points.
* Each year is a separate file with a slightly different cohort of corporations. Typically, there are c200 – 210 reporting each year, but only 158 corporations reported consistently for the full period 2014 to 2020.
* The pre-processed data is held on a series of seven excel tabs with c 200 rows and 88 columns and their transposed variants.

The **independent variable** data comprise an initial list of a dozen features of corporations that are thought to have a bearing on its compliance behaviour. Each of these potential IVs could be associated with a distinct causal mechanism that could influence a corporation’s compliance.

1. **Regulatory obligations**. There are two regulators in Peru with differing priorities with regard to governance and with different enforcement powers.
2. **Industry sector**. Each industry sector has different exposures to consumer and environmental activists. There are 10 sectors of note.
3. **Shareholding structure**. Minority shareholder rights is a key governance issue and so the presence of voting retail investors is key.
4. **Active local equity trading**. Exposure via the BVL to fundraising and active trading can be affected by governance. Firms are either active or not.
5. **Local bond issuance**. Bond investors are very interested in governance and probity. Firms either issue bonds locally or not.
6. **International equity listing**. International equity markets are tougher on governance than the local market. Firms either have an international listing or not.
7. **International bond issuance**. Bond investors are very interested in governance and probity. Firms either issue bonds internationally or not.
8. **Group affiliations/business model**. International subsidiaries or parents affects the process of governance. Firms may have four types of group affiliation.
9. **Supply chain relationships**. Globally branded suppliers or partners have an interest in the probity of their partners. Firms either have large globally branded partners or not.
10. **Scandals and SMV sanctions**. Sanctions in the form of fines and negative publicity over misdemeanours can impact ongoing profitability. Firms have experienced sanctions or a scandal or not.
11. **Multiple BVL listings**. The presence of affiliated firms in the listings could lead to efficiencies or higher standards. Multiple firms listed in the same group or not.
12. **Affiliations to major families**. Approximately twenty families dominate corporate ownership and politics in the country and could ha slack governance standards. Firms are either controlled by a leading family or not.

This data has been collected for one year, 2017, from a variety of sources according to the nature of the information for example, from corporations’ websites, from the SMV, from the BVL and from interviews with experts.

I consider the IVs to be relatively stable year on year, with only minor changes in the status of corporations and I judge this insufficient to justify recollection for the other 6 years of the study period.

Initial linear regression of overall population results explains c 40% of the variance in compliance between firms, the majority coming from based on just six of the features. The effect of features is cumulative - exposure to more features is associated with higher compliance:



However, these features do not appear to be independent – Interactions between features added significantly to the MLR adjusted R2 and csQCA analysis suggests some pairs of features provide sufficient causality for high compliance and though there is also equifinality – there are multiple ways in which features are associated with high compliance.

**Data pre-processing**

I have already begun to prepare the data:

Dependent variable:

* The “SI” and “NO” responses have been converted to “1” and “0”
* The small proportion of unavailable responses have been coded as “0” on the basis that a corporation wishing to assert its compliance would have done so but reasonably might leave a response blank if it did not comply.

Independent variables:

* Most of the features were binary categorical features and were represented as “1” and “0 on collection.
* Two features viz, Industry sector and Group structure/Business model have multiple categorical values and were represented by numerical dummy variables. In the case of industry sector, this was initially provided by the BVL but I found many errors and inconsistencies and so reformulated the sector classifications

Corporations:

* Although there were 208 corporations listed in 2017 which submitted compliance returns, four of these were in liquidation proceedings and so were removed from the analysis
* For inter-year analysis, because of ongoing listing and delisting, only 157 of the 204 corporations in 2017 reported throughout the period 2014 to 2020.

**The prediction task(s)**

Prediction can occur at three levels – 1) the degree of compliance with the code at the overall population level, 2) the level of compliance exhibited by an individual corporation, and 3) the compliance exhibited by an individual corporation with individual rules comprising the code. These require different treatment:

* Type 1 can be treated as a regression problem since although compliance can be expresses as an integer reflecting the discrete nature of the rules, the number of rules involved approximately equated to a numerical scale. Linear regression would be a good starting point.
* Type 3 is clearly a binary classification problem.
* Type 2 could be approached in different ways: it could be treated as an aggregation from Type 3 if compliance for all rules were to be predicted. It could also be modelled directly, either as a regression or as a classification problem if the numerical scale were transformed into a binary output for example above and below the mean compliance score.

For types 3) and 2) classifications, a wide array of models is appropriate as discussed later.

I intend to use training and validation protocols in developing models but will explore whether it makes sense to have both a validation and a test set. Given that there are only 200 cases it may be prudent to use cross validation.

**Classification of rules**

A crucial distinction between different types of rules that are included in the governance code has a bearing on the analysis and scope of modelling. Most institutional theory asserts that institutions are the rules of the game and leaves the matter there. A noted philosopher, J Searle distinguishes two key roles for rules which has a direct bearing on the governance market.

* Regulative rules serve to codify or order a pre-existing activity – most of the rules associated with car driving are of this form. People could be driving cars about without the Highway Code – albeit at some risk to them and others.
* Constitutive rules bring a new activity into being either by inventing the notion of the activity or by introducing new players into the game – the rules of chess are constitutive – people weren’t moving small figures around a chequered board before the game was invented.

These distinctions are crucial because some of the rules in the code introduce third parties into corporations’ governance process. This represents a major cultural change to governance which for many firms owned by a single extended family have regarded corporate governance as a private matter – even for publicly listed corporations with outside shareholders.

I have conducted preliminary analysis of the 88 rules in the code and identified c one half of the total which are constitutive in character where contestation is more likely and where model prediction is more crucial. On close scrutiny, I have highlighted 24 constitutive rules which are most instrumental in introducing third parties into the governance process of corporations or to changing the authority or engagement of existing third parties[[2]](#footnote-2).



Corporations complied overall with an average of 38.8% of the 24 constitutive rules compared with 71.7% of the 64 regulative rules (overall 62.5% of the 88 rules). The differences are clearly very distinct.

**Performance measurement and metrics**

Consistent with other classification tasks, it makes sense to use a confusion matrix to identify instances and proportions of correct and incorrect predictions as a basis for assessing prediction accuracy. Measures such as accuracy, precision and recall (sensitivity) will therefore be used, but unlike other use cases of classification models, there is only a certain degree of utility in achieving a high precision. This is because the underlying phenomenon is behavioural rather than existential. This means that false positives and false negatives are not just measures of the inadequacies of any model to predict an outcome, but they provide additional informational value. In particular:

* Type I errors or false positives are cases which have a low compliance despite having the features under examination thought to be driver of compliance. These firms would appear to be less sensitive to the features under consideration and provide policymakers the focussed task of encouraging higher compliance.
* Type II errors or false negatives are cases which have high compliance, but which do not exhibit the features included in the model. These firm provide an interesting opportunity to explore whether there are other factors causing these firms to have high compliance and provide opportunities features for further modelling.

In practice, three measures are likely to be prominent. Since the objective here is to predict positive outcomes, the Type II errors are an important issue – to what extent are positive cases falsely identified as negatives? The appropriate metric will be the True Positive Rate (TPR), Sensitivity or Recall which measures the proportion of predicted positives compared with the population. With a continued focus on true positives the positive Predictive Value (PPV), or Precision, will likely be useful as a measure of the ‘cleanness’ of the positive prediction itself. Finally, the ability to spot Type I errors will be considered and so involve the True Negative Rate (TNR), or Specificity.

The following chart provides a typology of metrics based on the confusion matrix, which emphasises the distinctions between a focus on either how reality is treated or on the quality of the prediction.

****

Source: Author’s preparation from publicly available information

**Scope management for the Capstone project.**

This is a large project for a 2- to 3-week engagement. I therefore suggest defining the scope of the modelling as described above to include a selection of constitutive rules and to predict their outcomes for the cohort of corporations on the BVL in 2017. This will provide a good test of ML approaches and also a basis for further work. My focus for the Capstone will therefore be as follow:

* The Independent variable will be a file of 204 records of 10 categorical features
* The dependent variable(s) will be the compliance results for 204 corporations for 24 constitutive rules which have received the greatest contestation
* The prediction task will be Type 3 i.e., to predict corporations’ compliance with one of the 24 constitutive rules.

**Modelling strategy**

With an emphasis on explainability, interpretability and usability, I intend to use a progressive strategy to find the optimal modelling approach – beginning with simple models and increasing complexity where justified.

1. Seek to simplify the features – it is likely that there are interdependencies in the features and for ML to be most effective, the feature list should be simplified. There are three ways to do this. Use of PCA to reduce the dimensions – this has the disadvantage of losing the connection with the original features and so losing explainability. Second, a lasso or ridge regression to highlight the features with most influence. Third, use of a logistic regression or decision tree which provide rankings of the most influential features, allowing the removal of the less impactful features on a second round. I will compare options two and three and use domain knowledge to shorten the feature list,

In addition to this review of the IVs, I will also check the dependent variables, ie the distinction between constitutive and regulative rules, using cluster analysis and make adjustments as appropriate.

1. With a reduced but more independent list of features, a first pass through relatively simple alternative classification model structures to explore overall levels of suitability and performance. Given the size of the case set, I will use train and test sets only. The initial models will include decision trees and logistic regression, and I will aim for a preliminary rating of suitability and performance on a common set of performance metrics likely derived from the confusion matrix.
2. A third pass on the more promising models to add sophistication, specific strategies and tuning with a mind to maintaining explainability and interpretability. The third pass models may well include logistic regression and random forest. I will then compare results before and after tuning to observe the impact
3. Consolidation of the most promising approach to refine and tune hyperparameters for optimal performance.

# **III Results**

**Initial exploration**

I began with a set of six classification models from scilearn:

* Logistic Regression': LogisticRegression()
* Support Vector Machines': LinearSVC()
* Decision Trees': DecisionTreeClassifier()
* Random Forest': RandomForestClassifier()
* Naive Bayes': GaussianNB()
* K-Nearest Neighbor': KNeighborsClassifier()

I used these initially in their basic form to ensure that they functioned with the data and to see whether, in their basic form, there might be obvious candidates for further examination.

Initial results for sensitivity are shown in the chart.

**A graph of lines with numbers and lines

Description automatically generated with medium confidence**

This chart shows sensitivity scores for the six classification models for each of the 24 constitutive rules under consideration. The rules are in rank order of compliance with the lower compliance rules to the right. Despite a degree of variance between models and across the rules where is clearly an association between predictive ability and the prevalence of less frequent positive cases.

No clear picture emerges from this collective analysis as to which might be the best model. In the next section,

**Optimizing the independent variables (features)**

The data on features had been collected on the basis that there could be a viable mechanism by which third parties might influence corporations, whether these be regulators, customers, business partners, shareholders or NGOs. However, there might well be some redundancy in this list because the features in the original list oof 10 might have interdependencies. This is crucial because classification algorithms assume feature independence. I have used three machine learning methods to reduce the number of features and to minimize any associations between them.

The first is Lasso regression which operates by identifying essentially redundant features. In addition, I have used logistic regression and decision tree s because both methods preserve the identity of features and provide information of their importance – in the case of logistic regression through coefficients, and in the case of decision trees through the attribution of a percentage share of contribution to the final classification.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Features | Lasso  top  features | Logistic  regression  coefficients | Decision tree  feature  importance | “Support” |
| int\_fin | 0.206050 | 1.072 | 25.3 | 3 |
| sector | 0.316850 | 0.951 | 28.8 | 3 |
| ret\_voters | 0.163480 | 0.821 | 8.1 | 3 |
| loc\_bonds | 0 | 0.497 | 0 | 1 |
| regulator | 0 | 0.428 | 0 | 1 |
| loc\_equity | 0 | -0.165 | 0 | 0 |
| multi\_list | -0.017316 | -0.229 | 15.2 | 2 |
| bus\_model | -0.011233 | -0.238 | 0 | 1 |
| blocksh | -0.060461 | -0.445 | 5.1 | 3 |
| family | -0.289507 | -1.193 | 17.6 | 3 |

There is some agreement between these classifications – particularly at the extremes and the distinction, where presented, between positive and negative influences. I will adopt an experimental approach starting with retaining the features with support from at least two of the three approaches, that is, to retain international finance, sector, retail voters, multi-listing, block shareholding and family, and add back other features if they make a rational improvement in results.

**Adjusting the classification thresholds for all models**

The threshold determines the manner in which classifiers make selections and this parameter can be adjusted from the default setting of 0.5. In this data set, there will be a preponderance of ‘negatives’ since compliance for constitutive rules is on average 39%. I expect to see results which favour the identification of negatives rather than positives, and so there is an argument to diverge from the norm. This is indeed the case, with initial results showing relatively high accuracy, specificity and precision but relatively low sensitivity (TPR).

There are four approaches: 1) ROC curve analysis to maximize the trade-off between TPR and FPR; 2) precision-recall curve analysis to identify the point at which precision and recall /sensitivity are balanced (the F1 score); 3) Youden’s J index – maximize “Sensitivity + Specificity – 1” ; and 4) F1 Score maximization – the harmonic mean of precision and recall (similar to 2).

For simplicity, I chose to go with option 3 in combination with the logistic classifier to explore the topic of the overall threshold. I ran 20 trials over multiple values of the threshold to find the one that optimizes the value of Sensitivity + Specificity.

Although the results of the logistic classifier are not very sensitive to the threshold value, s=a setting around 0.39 -0.4 would seem balance sensitivity and specificity.

A graph of a number of different colored lines

Description automatically generated with medium confidence

**Hyperparameter tuning**

In this section, I work with three of the models which identify features in the outputs, in view of the need for explainability. These were: logistic classification, a simple decision tree and random forest. I also chose a single important rule to focus on – rule 41

In the first instance I tuned these models with overall input tuning, then I used a simple grid search process using the algorithm GridSearch CV to identify hyper parameters relevant to each model. These are given in the table.

|  |  |  |
| --- | --- | --- |
| **Focus of tuning** | **Hyper parameters** | **Process** |
| Input data | Feature engineering | Use of Lasso and reference to feature rankings from Logistic and decision tree classifiers, from 10 to 6 features |
| Threshold for classifier models | Adjusted from 0.5 default to 0.39 as by maximizing sensitivity and specificity |
| Logistic | Model – liblinerar | GridSearchCV |
| Penalty – l2 |
| Regularization parameter C,  0.2682695795279725 |
| Decision Tree | Evaluation criterion |
| Maximum depth |
| Minimum sample size split |
| Random Forest | Number of estimators |
| Maximum depth |
| Minimum samples per split |
| Minimum samples per leaf |
| Maximum features |

I then compared results before and after tuning for each model, explore the potential improvement in metrics resulting from this initial hyperparameter tuning.

**Logistic classifier**

The logistic classifier showed good results for sensitivity but these were unchanged after tuning. What did change was the treatment of negatives or firms which did not comply.

Confusion matrices and metrics:

Before: [17, 13]

[ 5, 27]

Metrics:

* Accuracy: 0.71
* Sensitivity: 0.844
* Precision: 0.675
* Specificity: 0.567

After: [13, 17]

[ 5, 27]

Metrics:

* Accuracy: 0.645
* Sensitivity: 0.844
* Precision: 0.614
* Specificity: 0.433

It would seem on the basis of these confusion matrices, that the results are worse after tuning – the sensitivity is identical at 0.844, but the quality of the prediction is more diluted with precision moving from 0.675 to 0.614 – ie with additional false positives now included in the P call. This is apparent also in the graphic.

A graph of different colored bars

Description automatically generated

**Decision tree classifier**

This classifier showed similar initial results to the logistic classifier. However, tuning appears to have worsened.

Confusion matrices and metrics:

Before: [18, 12]

[ 6, 26]

Metrics:

* Accuracy: 0.71
* Sensitivity: 0.812
* Precision: 0.684
* Specificity: 0.6

After: [23, 7]

[10, 22]

Metrics:

* Accuracy: 0.726
* Sensitivity: 0.688
* Precision: 0.759
* Specificity: 0.767

The effect of this change on the metrics was to disfavour sensitivity from 0.812 to 0.688, and to favour accuracy, precision and specificity.

A graph with blue and orange bars

Description automatically generated

**The random forest classifier**

The before and after confusion matrices are not quite as good as above for basic sensitivity scores, but remained identical following tuning, indicating that the basic model had reached a stable solution and not further tuning had an impact.

Confusion matrices and metrics:

Before: [19, 11]

[ 8, 24]

Metrics:

* Accuracy: 0.694
* Sensitivity: 0.75
* Precision: 0.686
* Specificity: 0.633

After: [19, 11]

[ 8, 24]

Metrics : as ‘Before’

On this basis it would seem that the logistic classifier produced the most useful results though the tuning had the effect of worsening the Precision of the forecast while maintaining its Sensitivity.

**Significance testing**

I carried out two significance tests on the logistic classifier results to get a sense of how different might the predicted results be compared with a random selection. This was done using the logistic classifier on one rule – rule 41. – with the threshold at 0.39 and a reduced set of features – using the basic model.

A graph of a normal distribution

Description automatically generated with medium confidence

The chart shows the results from 1000 trials of a logistic classifier applied to rule 41 for accuracy scores compared with a random benchmark of 0.5. The results have mean 0.6704, standard deviation of 0.0552 and a p-value compared to the benchmark of 0.0007 indicating a significant degree of difference. I did not carry this out with the enhanced model, but would marginally more difference from the random base case.

A diagram of a logistic graph

Description automatically generated

The chart shows the results from 1000 trials of a logistic classifier applied to rule 41for sensitivity scores compared with two benchmarks of 0.5 and 0.39. The results have mean 0.679, standard deviation of 0.102, and p-values with respect to the benchmarks of 0.0197 and 0.0022 respectively indicating a reasonable degree of difference.

**Features ranking and Explainability**

Onre of the important considerations for this project is to be able to explain g=the colcusions and so models which provided such an out[put were favoures of=ver, say, SVM or PCA. However, although there was a reduced set of features used and this improved the classification metrics, there is not agreement over which features contributed to the output, as shown in the table (with the features in the same of=-rder as used above.

|  |  |  |  |
| --- | --- | --- | --- |
| Features | Post tuning feature ranking (1 to 6) | | |
| Logistic  Regression | Decision Tree | Random  Forest |
| int\_fin | 3 | 1 | 6 |
| sector | 2 | 2 | 1 |
| ret\_voters | 4 | 5 | 2 |
| multi\_list | 6 | 3 | 5 |
| blocksh | 5 | 6 | 4 |
| family | 1 | 4 | 3 |

Such a result would not build confidence in the process.

# **IV Conclusions**

The result above are only mildly encouraging to the extent that it does appear to be possible to predict the behaviour of corporation’s behaviour based on the collection of feature data that is publicly available.

However, I have two concerns. First, the effect of model tuning seemed to be to degrade the results rather than improve them. This might mean that the approach is flawed or that a more thorough approach is required. Second, the models that were tuned offered different rankings of the features determined as important and this would be be a concern for this type of modelling for what is perhaps a small database.

Although I have focussed on a single rule, the methodology is clearly transferable, though as the ensemble approach showed, predictability metrics decline as the incidence of positives declines for rules with reduced compliance.

Next steps would be to extend this analysis to the remaining rules and use these results to predict overall levels of compliance. Because of the likely variance effects, I would expect predictions of overall compliance to show less variance.

1. The headings are Conduct of the AGM, Conduct of the Board, Shareholders Rights, Risk Management and Transparency but other analysis has shown these categories not to be analytically meaningful since they each contain a bundle of rules which themselves are important. [↑](#footnote-ref-1)
2. In the modelling, I will check the distinction between Constitutive and Regulative rules via cluster analysis. [↑](#footnote-ref-2)