Predicting Cancellation of Power Plants using Machine Learning Methods

Xu Chen[1]

[1] Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton, NJ

ABSTRACT

- From 2005 to 2015, about 90GW of electricity generation capacity got cancelled annually due to various reasons. This leads to stranded assets, failures to meet electricity demand in developing countries, and errors in carbon dioxide emission calculation.
- Therefore, being able to predict whether a power plant is getting cancelled has multiple benefits. However, the cancelled power units comprise only 2% of total power generation units worldwide, making the classes extremely imbalanced when applying machine learning models.
- In this study, I apply various machine learning classifiers to predict the cancellation probability at power generation unit level. The problem of imbalanced class is tackled using multiple methods: resampling and generating synthetic samples, changing evaluation metrics, using penalized models, etc.
- I find that the Random Forest classifier achieve the best recall rate, and the gradient boosting classifier gives the best accuracy, precision, and F1 score among all the classifiers applied.

MOTIVATION

Can we predict which power plant unit will be cancelled?

- For investors, avoiding power plants that would be cancelled avoid sunk upfront costs and tranded assets.
- For policy makers, being able to detect power plants that are likely to be cancelled helps them prepared additional power generation units to avoid electricity generation shortage, especially in developing countries.
- To estimate and tackle outcomes with climate change, we want an accurate estimate of future committed carbon dioxide emissions to evaluate the potential damage.

DATA

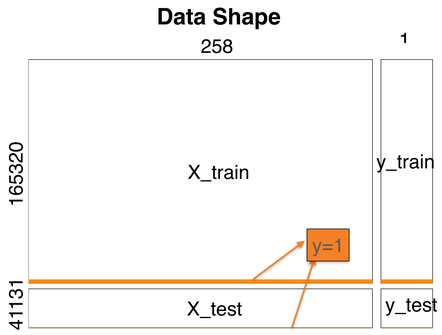
Data Preprocessing:

1. Data is from World Electric Power Plant Database from S&P Global Company, hereafter referred to as Platts database. This database documents all the power generation units in the world for each year.
2. I use Platts data from 2005 to 2016. Each year's data contains power generation unit's features. Each unit has a unique unit id. I use features and the target outcome (the class whether the power plant gets cancelled) from Platts.
3. I construct the target outcome to track whether the status of a power plant changed to cancelled during 2005 to 2016.
4. The features of each unit are selected from the latest year before cancellation. If the unit is not cancelled during this time period, information from the last year is selected as feature.
5. One-hot encoding was performed to transform categorical data to dummy variables.

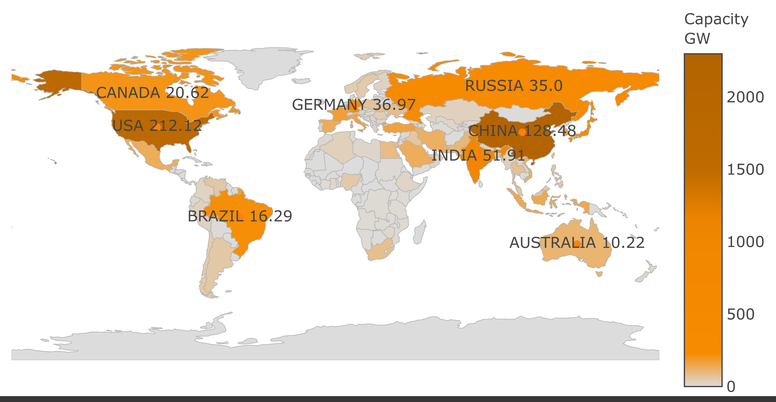
Data Description

- There are around 200,000 individual power plant units in 129 countries in the database.
- 1.7% are cancelled between 2005-2016, and are labeled y=1.
- Features include *business type*, *fuel type*, *country*, *capacity*, *status*, etc.

Data Shape



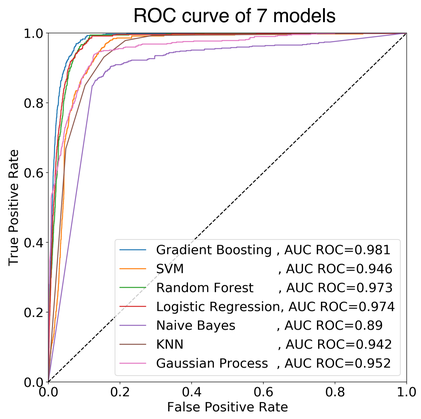
Cancelled Power Generation Capacity in Each Country from 2005 to 2016



APPROACHES TO IMBALANCED DATA

- **Applying resampling techniques**
  - Oversampling: upweight the minor class; advantage is that there is no information loss; used in logistic regression, random forest, and Naive Bayes classifiers
  - Undersampling: randomly downsample the major class; advantage is to reduce running time; used in Gaussian process, KNN, SVM, gradient boosting classifiers
  - Generating synthetic samples
- **Changing evaluation metrics from ROC-AUC score to F1 or recall**

Model Result

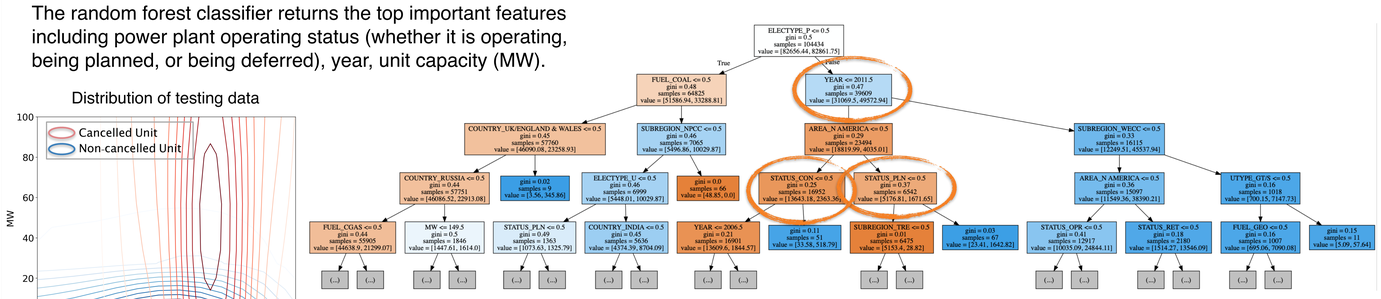


Classifier	Accuracy	Recall	Precision	F-1 score	ROC-AUC score
Logistic Regression	0.895	0.971	0.135	0.238	0.974
Random Forest	0.895	0.976	0.136	0.238	0.973
Gaussian Process Classifier	0.877	0.941	0.124	0.219	0.956
KNN	0.847	0.931	0.093	0.169	0.942
SVM	0.835	0.974	0.090	0.165	0.946
Naive Bayes	0.789	0.915	0.068	0.127	0.890
Gradient Boosting Classifier	0.926	0.961	0.179	0.302	0.981

Result of model performance using 5 metrics. Hyperparameters are optimized using grid search or random search with goal of maximizing recall. The best model are marked as orange under each metrics. Generally speaking, tree based models out-perform other models.

Model Interpretation

The random forest classifier returns the top important features including power plant operating status (whether it is operating, being planned, or being deferred), year, unit capacity (MW).



One decision tree in the random forest classifier, the important features in the decision tree are `YEAR`, `STATUS_CON` and `STATUS_PLN`

CONCLUSION

- I applied machine learning models including logistic regression, random forest, Gaussian process, KNN, SVM, Naive Bayes, and gradient boosting classifiers, and have succeeded in predicting cancellation of power plants with around 0.93 accuracy, 0.98 recall, 0.18 precision, 0.30 F1 score, and 0.98 ROC-AUC.
  - For high dimensional hyperparameter optimization, random search is used to speed up searching process.
  - Tree models out-perform linear models and probabilistic models.
  - High recall are achieved with low-precision trade-off.
- Probability of a power plant being cancelled is related to its operating status, generation capacity and expected year of commission.

FUTURE WORK

- Future work may use penalized models to approach the imbalanced sample, or apply more complex models such as neural network.
- This work not only predicts the cancellation probability of power generation units, but also helps identify the least likely plant to be cancelled that got cancellation, and the most likely plant to be cancelled that didn't (plants that beat the odds). This helps select least likely and most likely cases in future social science studies.

REFERENCES

1. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Pret- tenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Per- rot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 1000 Genomes Project Consortium, et al. A map of human genome variation from population-scale sequencing. Nature, 467(7319):1061–1073, 2010.
2. S&P Global Platts, World Electric Power Plant Database 2019 https://www.spglobal.com/platts/en