Final Project: Predicting the Cancellation of Power Plants Worldwide

Xu Chen
Woodrow Wilson School
xuc@princeton.edu

Abstract

Power plant cancellation leads to financial loss to investors and electricity shortage in developing countries. In this study I apply eight machine learning classifiers to predict the cancellation probability at power plant level. The problem of imbalanced class is tackled using multiple resampling methods and by changing the model tuning metric. High precision-recall trade off was observed. I find that the random forest classifier achieves the best recall rate, and the gradient boosting classifier gives the best accuracy, precision, and F1 score among all the classifiers applied. Tree models generally outperform the linear models and probabilistic models.

1 Introduction

Human activities related to energy use has contributed to climate change [1]. Power generation sector has contributed to 25% of global carbon dioxide emissions by 2010 [2]. To limit global warming to 2 degree Celcus by 2100 compared with pre-industrial age to avoid catastrophic climate impacts, the world's energy system needs timely transition. In line with the 2 degree Celcius target, power generation sector needs to decarbonize by mid-century [1]. Besides carbon dioxide emissions, electricity generation based on fossil fuel such as coal leads to local air pollution problem, which has negative impacts on human health and agriculture production [3]. In the meanwhile, developing countries are facing growing electricity demand and are trying to build new power plants to meet this demand.

However, controversies about carbon dioxide emissions and air pollution impacts are causing some of these newly built or newly planned power plants to be cancelled [4, 5]. This causes financial loss for investors and local governments (who are the investors of power plants in some countries), because the planning or construction of a power plant takes long preliminary investigation and large upfront cost. The cancellation of power plants also creates electricity shortage, especially for least developed countries who are relying on new power plants to meet the electricity demand.

Therefore, understanding why certain power plants get cancelled and being able to predict what kinds of power plants are more likely to be cancelled is beneficial. For investors, refraining from putting in large early on investment in these power plants avoid potential sunk cost. For policy makers, this information prepares them to have backup electricity generating units to resolve the electricity shortage when one or more power plant gets cancelled.

In this project, I use data from World Electric Power Plants Database (hereafter referred to as Platts database [6]) to predict whether a power generating unit gets cancelled. I apply machine learning classifiers including logistic regression (with l_2 penalty), support vector machine (SVM), Naive Bayes, random forest, K-nearest neighbors (KNN), Gaussian process classifier, gradient boosting classifier, and neural network to achieve binary classification. Features are taken from Platts database and include characteristics associated with the power plants such as fuel burned to generate

electricity, year in which the power plant was or is expected to be commissioned, plant status before cancelled, power generation capacity, etc.

This paper is organized as follows. Section 2 describes the related work in this field. Section 3 describes the dataset and feature engineering. Section 4 describes the methodology. In detail I discuss various machine learning models, how to handle imbalanced data and the choice of evaluation metrics. In section 5 I discuss the performance of different machine learning models, the outcome of different ways of handling imbalanced data, and the predictability of models. I discuss future work in section 6 and conclude with section 7.

2 Related Work

Recent reports found that power plant cancellations are becoming faster [4, 5]. Yet these studies cover a short time period and only compare between two or three years. Rigorous analysis of the cancellations of power plant is lacking, possibly due to the limitation of data coverage on power plants globally. Furthermore, in field of energy economics, machine learning techniques are mostly used to predict energy demand, electricity price, or crude oil price [7]. Application of machine learning techniques in other social science inquiries is still limited. Case studies on the cancellation of a power plant are usually based on empirical recognition or experience. For example, in Juraku et al. [8], the case of power plant was selected because the cancellation was "unprecedented".

This project tries to tackle the challenges in social science inquiries with the help of machine learning techniques. Besides the motivation mentioned in section 1, this study will also be beneficial for case selection in studies on power plant cancellation. Similar to the Fragile Family Challenge where we care about the children that beat the odds, this study will help identify power plants that we anticipate to highly likely be cancelled but are not, or power plants that we anticipate not to be cancelled but in fact are cancelled. Directing case studies to these "beating the odds" cases may reveal hidden factors such as corruption, local protest, or local community sentiment.

3 Data

3.1 Data Description

In this project, I use data from World Electric Power Plant Database from S&P Global Company (Platts database) [6]. Platts database documents all the power generating units worldwide since early 2000s. Every year, Platts publishes one version of database (one csv file for each year). I use Platts database from 2005 to 2016 where a unique *UNIT ID* is available for each power plant and is consistent across years, so that I can track plant's status change with time. Each year's Platts database contains about 200 thousands power units and 44 columns documenting information such as the power plant name, fuel type, generation capacity, generation status, power plant location, manufacture information, ownership information, etc in that specific year.

In Platts 2016, there are 214,009 power generating units worldwide adding up to 11,573GW of power generation capacity. Among them, 4,987 units were cancelled at some point in history, adding up to 1,277GW capacity. For the other generating units, 159,953 units (6,035GW) are operating, 4,494 units (723GW) are being constructed, 17,011 (2,154GW) are being planned. Other units are either retired, delayed, or deferred.

3.2 Data Preprocessing

The target variable is *CANCELLED*, where a power plant being cancelled corresponds to 1, and not cancelled correspond to 0. I construct this target variable by tracking the status change of each power plant from 2005 to 2016. When a plant was cancelled at some point, I label it as 1. When a plant has never been cancelled in history from 2005 to 2016, I label it as 0. A geographical distribution of the total capacity of cancelled units between 2005 and 2016 in each country is shown in Figure 1.

Feature variables are selected from Platts too. I found that columns in Platts that contain manufacture or ownership information are not accurate and not consistent with time, so I didn't use them as features. I include other features such as power plant location (country), generation capacity, fuel

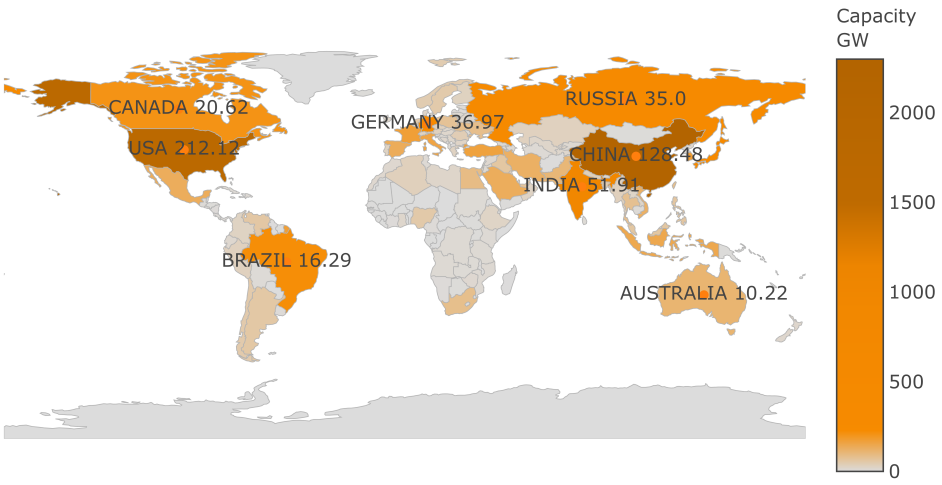

Figure 1: Cancelled Power Generation Capacity in Each Country from 2005 to 2016.

type, status in the year before cancelled, business type, etc. The features of each unit are taken from the latest year before the year when cancellation happens. For example, if a power plant was cancelled in 2016, feature variables take their values in 2015. If the unit was not cancelled during this time period (2005-2016), information from the last year (2016) is taken.

After constructing the target and feature variables, I split the data randomly into 80% training data and 20% test data.

3.3 Feature Engineering

By exploring and analyzing the database, I find that the number of cancelled power generation units varies from year to year. From 2008 to 2016 the cancellation experienced an increasing trend. Additionally, the number of cancelled units varies from country to country. This makes sense because each country has different power generation capacity and distinct policies on power sector development. Among the cancelled units, 75% of them were under planning phase before cancellation, 18% of them were already deferred before cancellation, only 1.7% units were operating before cancellation.

This gives me some insights when selecting features. I select features about the power plant's characteristics and that are possible to be related to whether a unit gets cancelled. A part of the feature names and what they represent are listed in Table. 1.

Table 1: Select features and description

Feature Name	Field Description	Data Type
YEAR	Year that the plant entered or is scheduled to enter commercial operation	int
BUSTYPE	The plant's parent company's business type	string
ELECTYPE	The plant's electricity production type, U(tility), P(rivate), etc.	char
COUNTRY	Country in which plant is located	string
FUEL	Primary fuel burned: COAL, GAS, etc.	string
UTYPE	Technology type of power plant: GT=gas turbine, HY=hydro, etc.	string
STATUS	Current unit status: CON=under construction, OPR=in operation, etc.	string
MW	Gross generating capacity of unit (MW)	float

I clean the data using the following procedures:

1. Drop the units located in countries that have never cancelled any power plant because the model will simply predict these units as not cancelled due to the lack of class 1 training data. After this step there are 129 countries left in the dataset.
2. For the *YEAR* variable, there are 16.6% missing values, indicating that expected year to start operating is unknown. Instead of filling NAs with mean, I fill NAs with 2025, which means the operation date is in the future. Platts's *YEAR* variable's largest value is 2024. So this method of filling NA is reasonable as we are assuming these plants will go into operation, just not in the near future where Platts has information about.
3. Categorical data is transformed into dummies by using one-hot encoding method. However, for some of the categorical features, they have huge amount of unique categories. Some of the category only appear once in the entire dataset. Simply applying one-hot encoding will enlarge the feature size and increase the sparsity of data. Thus I combine sparse categories (value counts less than 100 for feature *FUEL* and less than 1000 for feature *BUSTYPE*) and group them into another category: "Other", and perform one-hot encoding after.

After data cleaning, the feature size is (206651, 258), with 1.7% labelled with $y = 1$ and 98.3% labelled with $y = 0$.

4 Method

4.1 Machine Learning Models

Eight machine learning models are used to predict the cancellation of power generation units: logistic regression with l_2 panalty, support vector machine (SVM), Naive Bayes, random forest, k-nearest neighbors (KNN), Gaussian process classifier, gradient boosting classifier, and neural network. Logistic regression is chosen to be the baseline model, as it describe a linear relationship between dependent variables and independent variable.

Choosing these models is in favor of comparing the performance of linear models, tree models and probabilistic models. I expect linear models and tree models to have similar results because the majority of features are dummy variables, linear regression on 0s and 1s is equivalent with separation of 0s and 1s using tree model.

The Python machine learning packages I use is from SciKit Learn library [9].

4.2 Handling Imbalanced Data

Conventional machine learning algorithms often produce unsatisfying result when fed with imbalanced data. The reason is that most of the machine learning algorithms rely on minimizing the loss function, and the loss function is the sum of loss of each individual training sample. If the dataset is imbalanced, the minority class will have small contribution to the loss function, and the algorithm wouldn't properly classify minority class. In this paper, the impact of the ratio of imbalanced class will be discussed in section 5.1.

I use two different techniques to mitigate this imbalance issue: applying resampling techniques and changing evaluation metrics when tuning the models.

The main procedure of resampling is either increasing the frequency of minor class or decreasing the frequency of major class. This can be done by two techniques:

- Oversampling: upweight the minor class. The advantage of oversampling is that there is no information loss. It can be applied in logistic regression, random forest, and Naive Bayes classifiers. When applied to other classifiers, I found it to be too time consuming on a 2015 Macbook Pro machine with 16G RAM.
- Undersampling: randomly downsample the major class. The advantage of undersampling is to reduce running time, therefore it is best used in Gaussian process, KNN, SVM, gradient boosting classifiers, and neural network.

The result of two techniques are presented in section 5.2. They both solve the problem of imbalanced classes well.

I also tackle the imbalance issue by changing the evaluation metrics when tuning the hyperparameters. As discussed in section 1, we care about the power plant that got cancelled. Therefore, instead of using ROC-AUC score, I use recall as metric function in hyperparameter optimization. To reduce type-II error (false negatives) is the main interest in this study.

4.3 Hyperparameter Optimization

Hyperparameters are optimized using grid search method and random search method. Considering the training sample is large (around 200,000), grid search is performed in logistic regression and SVM because they have relatively low dimensional hyperparameter space and it is feasible to explore all the grids. Random search is used in random forest and gradient boosting because grid search takes long running time for these models. The advantage of using random search is that it is much more efficient compared to grid search, as it doesn't need to evaluate every single combination of the grid. Generally random search would achieve a better result because it has relatively higher "resolution" than manually selected grid [10].

Both grid search and random search use 3-fold cross validation and recall as the guiding evaluation metric.

4.4 Evaluation Metrics

Five metrics are used to evaluate model performance: accuracy, precision, recall, F1 score, and ROC-AUC score. I focus on recall, or true positive rate, as it represents the fraction of total positives that are correctly labeled as positive. In this case I'm interested in finding all power plant units that will be cancelled.

5 Results

5.1 Drawback of Imbalanced data

First of all, the drawback of imbalanced data is evaluated using the baseline model - logistic regression with l_2 penalty. For comparison, I resample the training dataset to have the same size but with three different class balance ratio. As a result, the first dataset has 1.7% of cancelled units, corresponding to the real world case. The second class has 30% of cancelled units. The third dataset has 50% of cancelled units, which represents perfect balance between two classes. The hyperparameter is optimized using grid search for all three datasets. The result is shown in Table. 2 and Fig. 2.

Table 2: Comparison of models under different ratios of imbalanced classes

dataset	hyperparameter	accuracy	recall	precision	F-1 score
1.7% cancelled units	$C = 46.4$	0.981	0.131	0.439	0.202
30% cancelled units	$C = 0.278$	0.915	0.954	0.169	0.288
50% cancelled units	$C = 0.278$	0.890	0.981	0.139	0.243

All three models are evaluated using the same testing data with imbalanced class. I find that slightly imbalanced training data (30% of y=1) does not affect the model performance too much, but highly imbalanced dataset leads to unsatisfying predictions. The result shows that failing to mitigate class imbalance causes the model to be useless because the model just labels the target variable with class 0. By applying resampling techniques to balance the class, the model may achieve lower accuracy, but has higher recall. Another interesting result is that when we use highly imbalanced training data, the model adds more l_2 penalty (large C value) on the coefficients to maximize recall.

5.2 Oversampling vs Undersampling

The section above highlights the importance of having a balanced training data. Therefore I apply oversampling or undersampling to the training data to achieve a balanced class ratio (1:1). In this section I compare between oversampling and undersampling and see whether they affect model

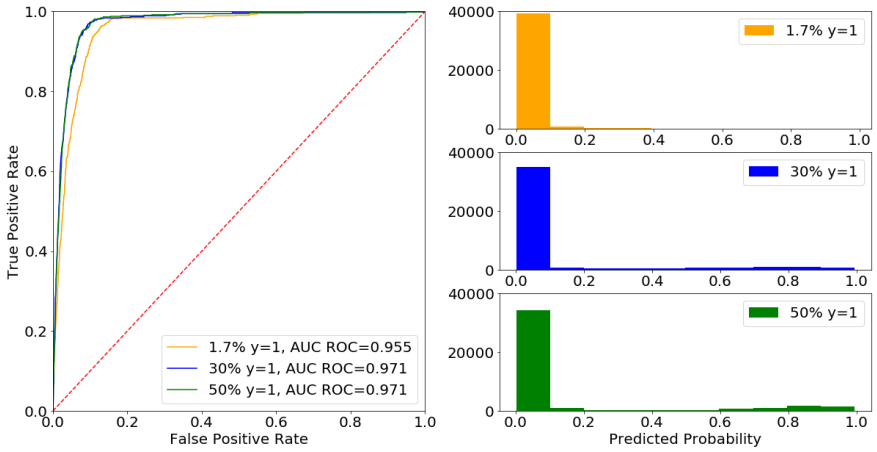

Figure 2: Performance of baseline model (logistic regreesion with l_2 penalty) under three datasets with different class ratios. The left figure shows ROC curves: model performance using dataset with 30% cancelled units and 50% cancelled unit are almost identical, and are generally better than using highly imbalanced data. The right panel shows the distribution of predicted probability on the test data. For the imbalanced dataset the model almost predicts all units to be not cancelled, which is a meaningless result.

results. I test these two resampling techniques using logistic regression, Naive Bayes and random forest. Other models are not tested because oversampling technique enlarges the data size and increase computer time cost greatly. The result is shown in Table. 3, with models tested using the same testing set.

Table 3: Comparison between oversampling and undersampling

Oversampling	Accuracy	Recall	Precision	F-1 score	ROC-AUC score
Logistic Regression	0.891	0.980	0.139	0.244	0.972
Random Forest	0.892	0.987	0.141	0.247	0.977
Naive Bayes	0.748	0.907	0.061	0.115	0.870
Undersampling	Accuracy	Recall	Precision	F-1 score	ROC-AUC score
Logistic Regression	0.890	0.981	0.139	0.243	0.971
Random Forest	0.901	0.985	0.152	0.264	0.981
Naive Bayes	0.836	0.869	0.088	0.160	0.877

All three models have similar performance using two resampling techniques. This is especially true for logistic regression. This makes sense because when large numbers of samples are drawn from the same population, the number of samples would not affect the linear relationship between dependent variables and independent variable very much. For Naive Bayes method, oversampling technique gives slightly better recall than undersampling, this may be due to the fact that Naive Bayes tries to learn the conditional probability of $P(y|x)$, and more samples lead to more accurate estimates.

This result illustrates that both oversampling and undersampling technique mitigates the issue of imbalanced training data, with minor difference on performance. For the sake of running time, I apply undersampling to all eight models in the following section for comparison.

5.3 Model Performance

Eight machine models are compared after applying the same undersampling technique to achieve 1:1 class ratio in the training data. The model performance is shown in Table. 4, evaluated using 5 metrics. The best result of each metrics is marked in red.

Table 4: Model performance

Classifier	Accuracy	Recall	Precision	F-1 score	ROC-AUC score
Logistic Regression	0.890	0.981	0.139	0.243	0.971
Random Forest	0.901	0.985	0.152	0.264	0.981
Gaussian Process Classifier	0.877	0.941	0.124	0.219	0.948
KNN	0.847	0.931	0.093	0.169	0.937
SVM	0.835	0.974	0.090	0.165	0.949
Naive Bayes	0.836	0.869	0.088	0.160	0.877
Gradient Boosting Classifier	0.926	0.961	0.179	0.302	0.979
Neural Network	0.774	0.726	0.052	0.097	0.843

The baseline model (logistic regression) gives a relatively good result: 0.981 recall and 0.971 ROC-AUC score. This indicates that features and outcomes have a good linear relationship. By looking at coefficients of logistic regression, I find that a power plant is more likely to be cancelled when it is deferred already or is still under planning phase. When a power plant is already operational, it is less likely to be cancelled. In Bangladesh or Ukraine, power plants are more likely to be cancelled, whereas in India, Bosnia-Herzegovina and Mexico, power plants are less likely to get cancelled. I also find that when a power plant is under government ownership, it is less likely to get cancelled. These results may give insights for future socioeconomic or political science studies.

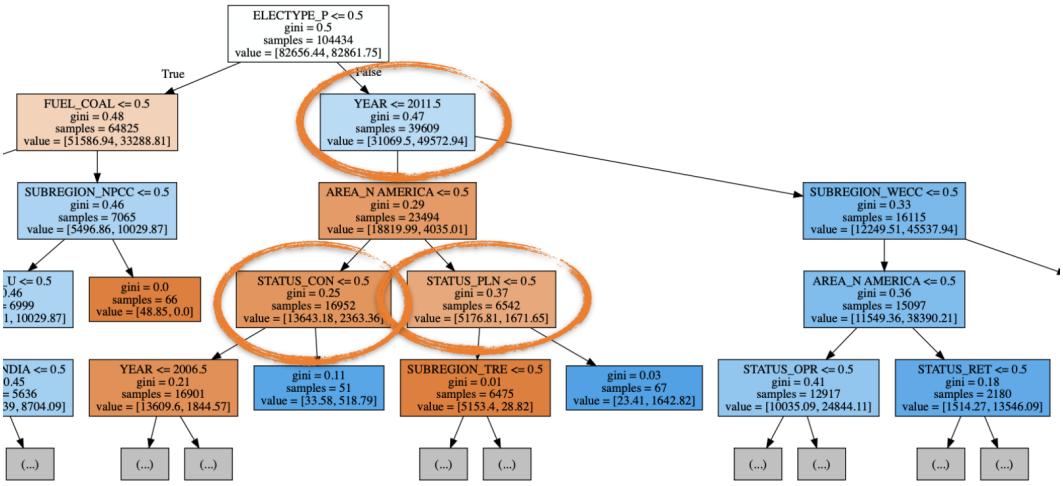

Figure 3: Part of a decision tree in the random forest model. Blue means the power plant unit is likely to be cancelled while orange means the power plant unit is not likely to be cancelled. This plot shows that year and unit status are important features.

Compared with the baseline model, two tree models (random forest and gradient boosting) have overall better results. This indicates that tree models capture the non-linear relationship in the data. Random forest returns the top important features. They include power plant operating status (whether it is operating, being planned, or being deferred), year, and unit capacity (MW). By visualizing a decision tree in the random forest classifier (Fig. 3), the result can be interpreted as follows: when a power plant has larger scheduled operation year or is being constructed or in the planning phase, it is more likely to be cancelled. This is consistent with what I find in logistic re-

gression. Additionally, the random forest classifier also indicates that power plants with larger unit capacity are more likely to get cancelled.

Neural network is trained using two layers with 100 neurons in each layer. The activation function is ReLu. The model gives a very conservative results: 66% of the predicted probability on test data falls in range [0.2, 0.8]. As a comparison, logistic regression only has 6.5% of predicted probability falls in that range, as shown in Fig. 2. The neural network model is very sensitive to the number of layers and the number of neurons in each layer. It is likely to be improved by better tuning the hyperparameters in future work.

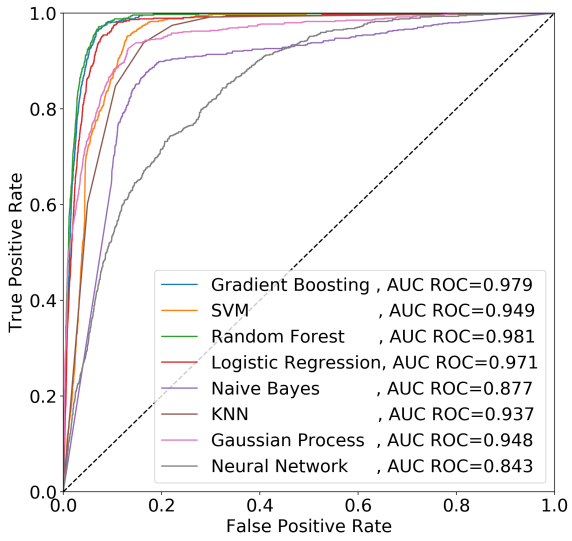

Figure 4: Performance of eight machine models. The baseline model (logistic regression) is shown in red.

Fig. 4 shows the ROC curve for all eight models. The baseline model is shown in red curve, as discussed above, only tree models out-perform the baseline model.

6 Future Work

To tackle the challenge of imbalanced training data, I applied oversampling and undersampling in this project. Other techniques to deal with the imbalanced class such as generating synthetic samples and using penalized models are out of the scope of this project. Future work may use these techniques to approach the imbalanced class issue.

Neural network was applied in this study as an initial attempt to try more complex models. Yet more work needs to be done to tune the neural network model better.

Time series analysis is also beyond the scope of this study. Ideally, if I am to predict whether a power plant will be cancelled the next year, I may need to treat each power plant as 11 samples from 2005 to 2016. For example, if a power plant was cancelled in 2016, this plant may count as 10 samples with class 0 from 2005 to 2015 and 1 sample with class 1 from 2015 to 2016. However I found that feature changes documented in Platts with time are often not due to actual changes, but due to the fact that Platts's previous information was wrong and was corrected the next year. For the rest of the cases, Platts have identical feature values between 2005 and 2015 with minor changes. In this problem, I'm already facing imbalanced sample with less than 2% of class 1. If I were to include all historical values to avoid forward looking issue, the fraction of class 1 would be even smaller. So in this problem I'm essentially predicting whether a plant is to be cancelled within the next 10 years or so. However further work focusing on time series analysis may be beneficial.

In terms of feature selection, I only used features from Platts database, which is another limitation on the data level besides the issue of imbalanced data. I used feature variable *COUNTRY* as a latent variable for economic and geopolitical indexes such as GDP per capita, GDP growth rate, political stability, electricity demand, power sector development policies, etc. I also used plant level features such as unit status and generation capacity. Yet for samples that have similar features, they still have different class labels. This creates a big challenge for the models because models cannot separate these samples based on their features. This can only be solved by collecting more feature data on the plant level, such as local sentiment towards the power plant, local ambient air pollution, and the relationship between power plant manufacturers and local community. Therefore this work highlights the importance of local survey data to further study the cancellation of power plants.

Given the results of this project, this study helps identify the least likely plant to get cancelled that however experienced cancellation, and the most likely plant I predict to be cancelled that actually didnt (plants that "beat the odds"). Similar to the Fragile Family Challenge, this study helps select the least likely cases in future social science inquiries. For example, if a plant is predicted highly likely to be cancelled however it is not cancelled, there may be corruption going on behind door pushing a risky plant going forward. If a plant is predicted not to be cancelled but in fact was cancelled, the plant may have experienced local protest, lack of funding, or other expected difficulties. In either case, further work may reveal hidden patterns that are not observed before.

7 Conclusion

In this paper, I applied machine learning models including logistic regression, random forest, Gaussian process, KNN, SVM, Naive Bayes, gradient boosting classifiers and neural network to predict the cancellation of power plants worldwide. Due to highly imbalanced class in the dataset, recall is used as metrics for model evaluation and hyperperameter optimization. Performance of two resampling techniques is studied. After resampling, the models predicted the cancellation of power plants satisfyingly, with around 0.93 accuracy, 0.99 recall, 0.18 precision, 0.30 F-1 score, and 0.98 ROC-AUC score. High recall is achieved with low precision trade-off. Among the eight models, tree models out-perform linear models and probabilistic models. Feature importance analysis shows that the probability of a power plant being cancelled is related to its operating status, generation capacity and expected year of commission. This work is beneficial to investors to avoid sunk cost in cancelled power plants, to policy makers to prepare backup electricity generating units, and to social scientist to conduct case studies on the ground.

Acknowledgments

I thank the Global Development Policy Center at Boston University for generously allowing me to use Platts database for this project.

References

[1] Rajendra K Pachauri, Myles R Allen, Vicente R Barros, John Broome, Wolfgang Cramer, Renate Christ, John A Church, Leon Clarke, Qin Dahe, Purnamita Dasgupta, et al. *Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change*. IPCC, 2014.

[2] Eric R Masanet. *Energy technology perspectives 2017: Catalysing energy technology transformations*. OECD, 2017.

[3] Drew Shindell, Johan CI Kuylenstierna, Elisabetta Vignati, Rita van Dingenen, Markus Amann, Zbigniew Klimont, Susan C Anenberg, Nicholas Muller, Greet Janssens-Maenhout, Frank Raes, et al. Simultaneously mitigating near-term climate change and improving human health and food security. *Science*, 335(6065):183–189, 2012.

[4] Endcoal, global new coal plant pipeline keeps shrinking. `https://endcoal.org/2018/03/global-new-coal-plant-pipeline-keeps-shrinking/`. Accessed: 2019-05-14.

[5] Ieefa update: India coal plant cancellations are coming faster than expected. `http://ieefa.org/india-coal-plant-cancellations-are-coming-faster-than-expected/`. Accessed: 2019-05-14.

[6] S&p global platts, world electric power plants database. `https://www.spglobal.com/platts/en/products-services/electric-power/world-electric-power-plants-database/`. Accessed: 2019-05-14.

[7] Germán G Creamer, Hamed Ghoddusi, and Nima Rafizadeh. Machine learning in energy economics and finance: A review. *Available at SSRN 3270251*, 2018.

[8] Kohta Juraku, Tatsujiro Suzuki, and Osamu Sakura. Social decision-making processes in local contexts: an sts case study on nuclear power plant siting in japan. *East Asian Science, Technology and Society: an International Journal*, 1(1):53–75, 2007.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[10] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.