Risky Business: Predicting Cancellations in Imbalanced Multi-Classification Settings

Anand Deshmukh, Meena Kewlani, Yash Ambegaokar, Matthew A. Lanham

Purdue University Krannert School of Management

deshmuk6@purdue.edu; mkewlani@purdue.edu; yambegao@purdue.edu; lanhamm@purdue.edu

Abstract

We identify a rare event of a customer reneging on a signed agreement, which is akin to problems such as fraud detection, diagnosis of rare diseases, etc. where there is a high cost of misclassification. Our approach can be used in all cases where the class to be predicted is highly under-represented in the data (i.e. data is imbalanced) because it is rare by design; there is a clear benefit attached to this class' accurate classification and even higher cost attached to its misclassification. Pre-emptive classification of churn, contract cancellations, identification of at-risk youths in a community, etc. are potential situations where our model development and evaluation approach can be used to better classify the rare but important events.

We use Random Forest and Gradient Boosting classifiers to predict customers as members of a highly underrepresented class and handle imbalanced data using techniques such as SMOTE, class-weights, and a combination of both. Finally, we compare cost-based multi-class classification models by measuring the dollar value of potential lost revenue and costs that our client can save by using our model to identify at-risk projects and proactively engaging with such customers.

While most research deals with binary classification problems when handling imbalanced datasets, our case is a multi-classification problem, which adds another layer of intricacy.

Introduction

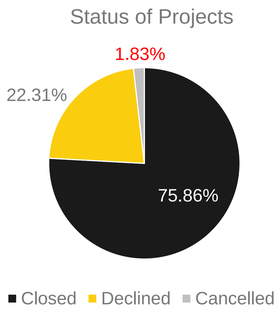
Identifying new sales opportunities and allocating resources against the best potential-revenue generating accounts is a challenging problem companies face. When a customer reneges on a prior commitment, companies not only bear the loss of potential revenue, but also the sunk costs associated with acquiring the customer's business.

Our industry partner is in the business of fixing/improving a product owned by their customer. Acquiring the customers is a very involved and resource-intensive process and once the customer signs the contract, several internal and external resources are employed in the planning and execution of these projects. Hence, such unforeseen cancellations pose a significant risk to our industry partner.

We study the use of machine learning techniques in predicting if a customer might cancel a deal after initially agreeing, and build a model to identify at-risk projects, thereby providing our industry partner the decision-support required to proactively engage with the customer and save their business.

We use classification techniques to classify a project as:
1. Closed: Project is successfully completed
2. Declined: Project is declined by management
3. Cancelled: Customer reneges on the contract

As visible from the graph alongside, there is heavy imbalance between the classes. We treat this using the following techniques:
1. Resampling
2. Class weights
3. Combination of both



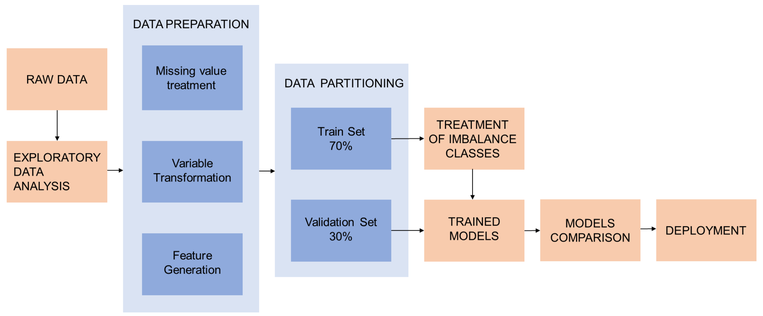
Data

Data Sources

1. Industry Partner: Our industry partner provided attributes of all the projects undertaken in the past 12 months, with over 300,000 observations. It included information regarding the projects, customers, the product being fixed, leads, lead sources, industry partner's employees, representatives that are involved in the project, and many more. The status of the projects (Closed/Declined/Cancelled) is the response variable.

2. Publicly available (zip code level) demographic data about income, educational levels, unemployment rates, and population were used to create clusters of zip codes.

Methodology



Feature Engineering
1. Discount ratio: Discounted price offered versus market price
2. Clusters of zip codes using K-means algorithm
3. Features created to quantify the experience and win-rate of salesforce

Interesting finding from exploratory data analysis: Tougher the customer, greater the discounts (refer to graph).

After feature engineering, EDA, and preprocessing the data, we treat the class imbalance on the train set.

Model Building and Comparison/Selection

We use Random Forest and Gradient Boosting classifiers for this multi-class classification problem.

Treatment of Class Imbalance
1. Resampling using Synthetic Minority Over-Sampling Technique (SMOTE):
   i. Auto: Over-sample all classes to match the majority class
   ii. Minority: Over-sample the minority class to match the majority class
   iii. Custom over-sampling using dictionary as an argument
2. Class weights

Since the classes are imbalanced, we use the following **Evaluation Metrics:**

Evaluation Metric	Formula	Interpretation
Recall	$\frac{(True\ Positive)}{(True\ Positive + False\ Negative)}$	When an instance actually falls within a class, how often does the model correctly classify it as falling in this class
Precision	$\frac{(True\ Positive)}{(True\ Positive + False\ Positive)}$	When the model predicts an instance to fall within a class, how often does it actually fall within the class

Additionally, we calculate the costs saved by our industry partner post running our model.

Cost Matrix ($C_{i\,x\,j}$):

		Predicted Status		
		Cancelled	Closed	Declined
Actual Status	Cancelled	$3,700	-$500	$0
	Closed	-$500	$0	$0
	Declined	-$500	-$40	$40

Assumption: 10% of projects that would have been cancelled can get saved.

Confusion Matrix for each model (3×3): $CF_{i\,x\,j}$

Cost Saving Per Project = $\sum_{i=1}^{3}\sum_{j=1}^{3}\left(\frac{(C_{i\,x\,j}) \times (CF_{i\,x\,j})}{\textbf{\textit{Number of Projects}}\ (N)}\right)$

Finally, we pick a model that has the best performing evaluation metrics across the three classes and enables our industry partner to save the highest potential revenue and cost by correctly classifying the projects that would close, get declined by management, or get cancelled by customers.

Results

Classification Technique	SMOTE	Class Weight	Cost Saving Per Project
Random Forest	None	None	$30.99
Random Forest	Auto	None	$32.22
Random Forest	Minority	None	$29.53
Random Forest	Custom	None	$31.97
Random Forest	None	Balanced	$32.39
Random Forest	None	Custom	$32.22
Random Forest	Auto	Balanced	$32.61
Random Forest	Auto	Custom	$32.60
Gradient Boosting	None	None	$35.44
Gradient Boosting	Auto	None	$32.44
Gradient Boosting	Custom	None	$34.09

The table above illustrates the combinations of models we ran for the purpose of this study. For each model, we calculated the cost savings per project based on the formula mentioned earlier. We observe that while SMOTE and class weights increased the cost-savings for the Random Forest Classifier (in isolation as well as in unison), the base model of the Gradient Boosting classifier outperformed all the models and performed better without the treatment of class imbalance.

On the basis of the cost savings, the top 3 models (highlighted above) were:

1. *Gradient Boosting (Base model)*

Classification Technique	Classes	SMOTE	Class Weight	Precision	Recall	F1
Gradient Boosting	Cancelled	None	None	0.96	0.58	0.72
	Closed			0.84	0.95	0.89
	Declined			0.69	0.42	0.52

2. *Gradient Boosting (with Custom SMOTE)*

Classification Technique	Classes	SMOTE	Class Weight	Precision	Recall	F1
Gradient Boosting	Cancelled	Custom	None	0.96	0.56	0.71
	Closed			0.84	0.94	0.89
	Declined			0.66	0.40	0.50

3. *Random Forest (with Auto SMOTE and Balanced Class Weights)*

Classification Technique	Classes	SMOTE	Class Weight	Precision	Recall	F1
Random Forest	Cancelled	Auto	Balanced	0.85	0.58	0.69
	Closed			0.81	0.87	0.84
	Declined			0.42	0.31	0.36

Best Performing Model

Classification Technique	SMOTE	Class Weight	Annual Cost Saving
Gradient Boosting	None	None	*$10.63 Million*

Conclusions

1. We develop a model to predict if a project undertaken by our industry partner would successfully get completed, get declined by the company, or if the customer would renege on the contract and cancel the project
2. We use a Random Forest classifier and a Gradient Boosting classifier for this multi-class classification problem. The imbalance in classes is treated using SMOTE, setting class weights, and a combination of the two.
3. Models are evaluated by comparing the potential revenue and costs they save as well as the precision and recall scores of predictions. The precision and recall scores of the highest cost saving model is also the highest amongst the models developed.
4. By deploying our best performing model, our industry partner can **save $10.63 million annually**.

Acknowledgements

We thank our industry partner for sharing their business problem and data with us. The Purdue BIAC partially funded this work.