RESEARCH

Quantifying decision making for data science: From data acquisition to modeling

Saurabh Nagrecha[1] and Nitesh V. Chawla[1*]

*Correspondence: nchawla@nd.edu
[1]iCeNSA, Department of Computer Science and Engg., University of Notre Dame, Notre Dame, IN 46556, USA
Full list of author information is available at the end of the article

1 Relative Costs of Acquisition and Model Development

In practice, price points for acquisition of external data are known, since these are to be negotiated with the external data provider. Cost-to-company of a data science development team to make sophisticated models can be determined as well. Since the derivation of these values is outside the scope of this paper, we assume a set of values for each of data acquisition and modeling component for illustrative purposes. These values can be swapped out with the appropriate values in a real use-case scenario. Our framework can also allow for conducting scenario analyses on different cost functions.

The NPV cost of model development can be directly added to the NPV cost of prediction errors, since it can be assumed that the same cost C_{Model} gets added to every batch of prediction tasks. Mathematically, we are just linearly decomposing the Total Cost from Main Paper Subsection 3.3 to its constituent components. From an accounting perspective the investment into model development would include salaries and benefits for the team, data storage and management costs, compute infrastructure, etc. The returns component of this is not so cut-and-dry. Returns are not necessarily immediate, depend on potential for the model to be improved, and are subject to externalities.

2 Classification Techniques

As a dual to the cases in Main Paper Subsection 3.1, we use the following classification techniques: Decision Tree (`DT`), Cost Sensitive Decision Tree (`CSDT`), CS-DT with Dynamic Cost Matrix (`CSDT-dyn`). `CSDT` uses a static cost matrix, whereas its dynamic counterpart `CSDT-dyn` has a cost matrix that is dependent on NPV. Detailed discussion on these particular costs is in Main Paper Subsection 3.3. To simulate an increasing complexity of the model used, we compare varying degrees of model complexity in the underlying machine learning model.

Main Paper Subsection 3.6 identifies the key classifiers being considered. The implementations for each classifier used in this paper are as below:

- `DT`: C4.5 classifier [1]
- `CSDT`: A MetaCost wrapper around `DT` [2]
- `CSDT-dyn`: A MetaCost wrapper around `DT`, but with varying cost matrix

Motivated by the above concept of dynamic cost matrices, we introduce the notion of a classifier which uses CM_{mod} as the cost matrix built into its training process. We call this classifier scheme CSDT-dyn since it is a dynamic cost sensitive decision tree.

3 Discount Rate

An important parameter for NPV is the Discount rate. It is an indicator of opportunity cost of the investment that went into acquiring external data or model development. Upon increasing this discount rate, we implicitly say that $X is now worth less in the future than it was before. The discount rate is a direct (monotonic) indicator of how severely one devaluates costs in the future. The monotonicity of this argument emphasizes the value of *immediate gains*. An intuitive, but simplistic, picture of this idea would be that a strategy which makes $X worth of misclassification errors in $t = t'$ is preferable over one which makes $X worth misclassification errors at some $t'' < t'$ (for a constant discount rate).

4 Experimental Role of Cost Factor

In Main Paper Subsection 3.3, we use the idea of heterogeneous costs of false negatives to false positives, which we call η. This is an important parameter in imbalanced class learning. A classifier trained with low η in mind would permit more false negatives than one with a high value of η. The tweaking of this parameter empowers the user to directly plug in their cost model for false negatives versus false positives. In case of variable cost models, one can use the approach proposed in Main Paper Subsection 3.5 to obtain an estimate of η from the training data.

5 Experimental Data Details

UCI data Simulating External Data To simulate "external data" in a dataset where we know all the feature and instance information, we perform a hold-off in the instance space or the feature space or both as applicable. The held-off data is reintroduced as "external data", available at a price. To mitigate any statistical sampling and partitioning bias, this is repeated for multiple iterations with different hold-offs and the entire set of experiments is repeated with a shuffled instance space. The amount of data reintroduced from the hold-off demonstrates how much external data one needs to invest in. This enables us to compare not only singular strategies, but also hybrid strategies involving model development of a certain degree in conjunction with a certain model of external data.

Medicare Data The following delineation was used in order to separate doctors from other health care professionals:

- Doctor: MD, DDS, DDM, DPM, PSY, PT, DO, OD, MNT, DC, and CP
- others: OT, SCW, AU, AA, PA, CSW, CNS, CNA, LPN, LVN, RN, BSN, MSN, CRNA, CNM, COHN, NP, and NR

In order to eliminate features which would be directly indicative of whether the professional is a doctor or not, we removed the school name and specialization fields from the data. For example, if a certain provider's Primary Specialization is "General Surgery", it is a direct indicator that they are a doctor.

Open City Data At any given time, the most recent data on the website is about seven days old. This dataset is manually typed into the CLEAR (Citizen Law Enforcement Analysis and Reporting) system by the Chicago Police Department. We have 250,000 instances of crime reports which span over 5 years (2010-14). We predict in increments across years in order to batch our results. The dataset contains a

rich array of features about each reported incident. Relevant to the prediction problem are features like date, time, location and type of incident. The data also contains a record of softer details like description of crime, location and the circumstances surrounding it.

The external dataset has features including the percent of housing that is crowded, households that are below poverty, individuals at or above the age of 16 who are unemployed, individuals younger than 18 or older than 64, per capita incomes, and a "hardship index".

6 NPV Evaluation for External Data Strategies

Here we list the full tabular dollar values of the external instance strategies when applied to the various datasets in this paper.

Pendigits Dataset First we consider the pendigits dataset for which (by simulation), both Case B and C are possible in terms of external data. Since the data has been simulated, we can see a sweep in terms of how much external data can/should be used. This is not the case for the rest of the datasets, where the external data is in the form of external features.

From Table 1, we see that the baseline costs are only a function of the classification technique used. This is to be expected since the baseline is calculated on what the NPV costs are without any external data. As the cost of external data increases, it becomes less feasible to make a misprediction, especially a false negative and so the trend in each row is

Medicare Data For the Medicare data, only the external transactional data is available. This is an example of timed batches with queries for external feature data, making it an example of Case D.

Open City Data Lastly, Open City Data uses external feature data from an aggregated set of indicators made available to it in one initial dump. This is still an example of Case D external data, but the insights from it are very different from those in the Medicare data.

7 More on Feasibility

For a given cost factor, feasibility is a convex-bound function of external data costs and model development costs, i.e. if external data at $\$X$ is infeasible, then for the same set of parameters, external data at a price greater than $\$X+\Delta X$ is guaranteed to be infeasible.

On the contrary, it should be noted that factors such as predictive performance and cost factors do not offer convex guarantees in optimizing NPV costs. This means that 1) increasing increasing the cost factor, and/or 2) adding more external data instances does not guarantee lower NPVs of costs. Both of these are confirmed in Figures 6 and 8 from the Main Paper.

8 The special case of free external data

In the special case of free external data instances in Figure 3, the simple decision tree can still turn out to be comparatively infeasible if too few external instances are introduced. This can solely be attributed to the NPV of prediction error costs being higher than baseline, since the data costs are factored out.

In the case of Medicare data, free external data would be helpful in formulating a greatly feasible strategy. More importantly, the model development scenario leading up to a dynamic cost matrix decision tree outweighs the free external data scenario.

For Open City Data, free external data immediately pays off by brushing roughly $40,000 off the baseline NPV. However, it shows its usefulness when used in conjunction with model development.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
SN and NVC conceived of and designed the research. SN implemented the empirical analysis. SN and NVC wrote the paper.

Acknowledgements
This material is based upon work supported by the National Science Foundation under Grant No. 1447795.

Author details
[1]iCeNSA, Department of Computer Science and Engg., University of Notre Dame, Notre Dame, IN 46556, USA. [2] , , , .

References

1. Quinlan, J.R.: Bagging, boosting, and c4. 5. In: AAAI/IAAI, Vol. 1, pp. 725–730 (1996)
2. Domingos, P.: Metacost: A general method for making classifiers cost-sensitive. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 155–164 (1999). ACM

Tables

Table 1 Net Present Values for External Instance Strategies. For the `pendigits` dataset, various NPVs for each scenario are enumerated here. Case B is for transactional batch-wise instances of external data, whereas Case C represents equivalent one-time dumps of external instances of external data.

Ext Data Strategy	Technique	Fraction external	Baseline Costs	External Data Costs (per instance)								
				0	0.01	0.05	0.25	0.5	1	5	10	50
B	DT	0.2	1453.611	1705.34	1705.34	1712.48	1748.21	1791.07	1883.95	2598.41	3498.63	10678.92
		0.5		1702.43	1702.43	1723.86	1809.59	1923.91	2145.39	3945.82	6189.21	24143.51
		1		1151.26	1158.41	1194.13	1372.75	1594.23	2044.34	5638.05	10124.84	46033.44
		2		1224.12	1238.41	1309.86	1667.09	2117.19	3017.41	10197.70	19171.28	90988.46
		5		851.03	893.90	1072.51	1972.73	3094.42	5337.82	23292.11	45733.20	225269.03
	CSDT	0.2	844.639	1033.60	1033.60	1040.74	1076.47	1119.33	1212.21	1926.67	2826.89	10007.18
		0.5		498.40	498.40	519.83	605.57	719.88	941.36	2741.79	4985.19	22939.48
		1		698.66	705.80	741.53	920.14	1141.62	1591.73	5185.45	9672.23	45580.83
		2		659.30	673.59	745.04	1102.27	1552.37	2452.59	9632.88	18606.45	90423.64
		5		135.89	178.76	357.37	1257.59	2379.29	4622.68	22576.98	45018.06	224553.89
C	DT	0.2	1453.611	1553.25	1555.25	1566.25	1622.25	1691.25	1829.25	2935.25	4317.25	15374.25
		0.5		1161.39	1167.39	1195.39	1333.39	1506.39	1852.39	4616.39	8071.39	35713.39
		1		1029.03	1042.03	1098.03	1374.03	1720.03	2411.03	7939.03	14850.03	70134.03
		2		1226.22	1253.22	1364.22	1917.22	2608.22	3990.22	15047.22	28868.22	139436.22
		5		512.62	581.62	857.62	2239.62	3967.62	7422.62	35064.62	69617.62	346037.62
	CSDT	0.2	844.639	440.50	442.50	453.50	509.50	578.50	716.50	1822.50	3204.50	14261.50
		0.5		371.45	377.45	405.45	543.45	716.45	1062.45	3826.45	7281.45	34923.45
		1		598.02	611.02	667.02	943.02	1289.02	1980.02	7508.02	14419.02	69703.02
		2		529.19	556.19	667.19	1220.19	1911.19	3293.19	14350.19	28171.19	138739.19
		5		634.29	703.29	979.29	2361.29	4089.29	7544.29	35186.29	69739.29	346159.29

Table 2 Net Present Values for Medicare data. This dataset is an ideal example of developing an in-house model instead of acquiring external data.

Technique	In-house NPV	External Data Cost (per query)					
		0	0.1	0.5	1	5	10
DT	46,435.26	7,275.04	7,886.83	10,333.97	13,392.89	37,864.29	68,453.54
CSDT	5,754.58	8,412.35	9,024.14	11,471.28	14,530.20	39,001.60	69,590.85
CSDT-dyn	492.44	7,632.76	8,244.55	10,691.69	13,750.61	38,222.01	68,811.26

Table 3 Net Present Values for the Open City Data. This dataset clearly benefits from adding external data.

Technique	In-house NPV	External Data Cost (per query)				
		0	0.01	0.1	0.25	0.5
DT	189,050.15	149,642.63	151,538.02	168,596.56	197,027.46	244,412.30
CSDT	92,878.37	73,888.21	75,783.60	92,842.14	121,273.04	168,657.88
CSDT-dyn	98,102.80	73,882.00	75,777.39	92,835.93	121,266.84	168,651.67