Applying AIML methods to Portuguese Bank Data Set

UCB AIML Module 17

Mike Jones

11.11.24

EXECUTIVE SUMMARY

Artificial Intelligence and Machine Learning models can be applied to business data sets for the purpose of more clearly understanding a business's operations and increasing the efficiency, productivity and quality of it's processes. Adopting and applying these innovations give competitive advantage to the organizations who invest in these capabilities.

For example, consider the Portuguese Banking Data Set (PBDS), which represents a 1.5 year marketing campaign to enlist new customers for their new bank deposit product offering. The bank wants to achieve the business objective of signing up customers for deposits. By using AIML methods to predict which marketing/sales calls will convert sales leads into paying customers, they can accurately forecast and prioritize phone calls with a high likelihood of conversion. Using AIML classification models and methods, sales leads can be profiled and classified into one of two predictive categories:

1 "likely to convert lead to a sale", or
0 "unlikely to convert".

This has business value as it allows sales agents to forecast and prioritize whom they will call from their lists of sales leads. Thus, marketing and sales activities can be made more efficient than simply calling people from leads lists, without any further business intelligence. The discussion below covers how to apply AIML modeling techniques to the PBDS. Additionally, the CRISP-DM (Cross Industry Standard Process for Data Mining) is illustrated as the way of organizing and achieving the result.

INTRODUCTION AND BACKGROUND

During the 2008 global financial crisis, the institutional investment landscape rapidly shifted away from a state of balance and equilibrium into a period of uncertainty and chaos. During the rapidly changing events of this era, most large organizations were immediately inflicted with severe negative consequences that had a lasting strategic impact.

In order to regain balance, European banks brought to market a whole new set of product offerings. These represented new investment opportunities for creditworthy customers. In this way, institutions could quickly raise capital for the purpose of achieving stability in the course of reformulating and establishing a new coherent financial situation.

In one such case, the product took the form of bank-deposits with attractive interest rates. The Portuguese Bank Data Set examined in the course of this exercise represents a summary of descriptive statistics for that product's marketing campaign during this era. Prospective customers were contacted by the bank's agents via telephone in a marketing campaign whereby deposits where promoted and sold. Roughly 41,000 leads were contacted and 4,600 subscriptions sold (~11% conversion).

The Portuguese Banking data set was collected over year and half. It was previously analyzed and the findings written up in the paper: ***Using Data Mining for Bank Direct Marketing: An application of CRISP-DM Methodology*** by Sergio Moro, Raul Laureano & Paulo Cortez.

The paper uses the CRISP-DM methodology as a procedural structure for establishing and refining it's Predictive Analytics / Business Intelligence results. The authors use three Statistical Machine Learning models: Naive Bayes, Decision Trees, and Support Vector Machines (NB, DT & SVM, respectively). First they establish an initial result and then make iterative improvements to increase the model's efficacy. Model effectiveness is evaluated using ROC_AUC & LIFT_AUC to drive model refinements in successive iterations, building toward final measures of:

	ROC	LIFT
NB	0.87	0.82
DT	0.86	0.79
SVM	0.93	0.88

The current task works toward 3 goals:

```
[1]  Replicate, build upon, and extend findings in the paper,
[2]  Establish a foundation for testing and comparing models, and
[3]  Examine, compare and apply four classification models:
[3a] LGRG - Logistic Regression,
[3b] DT   - Decision Trees,
[3c] KNN  - K-Nearest Neighbors, and
[3d] SVM  - Support Vector Machines.
```

DETAILS

See EXHIBIT_1 section [1] (E1.1) CRISP-DM - Data Mining Six Phase Cycle:

```
[1a] Business Understanding
[1b] Data Understanding
[1c] Data Preparation
[1d] Modeling
[1e] Evaluation
[1f] Deployment
```

The outline starts with CRISP-DM and uses it as a guide for both the implementation of the code itself as well as the structure of this paper. The outline gives the basic order in which the data mining task (aka AIML) analysis is performed. Note that I'm using the three concepts: Data Mining, Artificial Intelligence and Machine Learning, as interchangeable terms here. The same methods of regression and classification are used in each of these disciplines, which share so many significant similarities to the point of being indistinguishable in many cases.

[1] BUSINESS UNDERSTANDING

One fundamental idea in business is the basic Accounting Equation (E1.2):

$$A = L + E$$

where a general idea is to increase Owner Equity by increasing Assets or decreasing Liabilities. Another fundamental related concept is Gross Profit, given as (E1.3):

$$GP = R - C$$

where revenue minus costs is used to determine the profit for a business activity.

AIML methods can be used in a Marketing and Sales context in order to predict whether a Lead will convert into paying Customer. This information can be used to make Marketing and Sales efforts more efficient and profitable. Thus, AIML serves to optimize sales throughput by increasing Revenue while decreasing Costs. Using AIML models to predict which calls are likely to lead to sales conversions is good way to forecast and prioritize calls which. The Predicted Positive calls (i.e. those having a high likelihood of converting (TP/PP - PPV)) versus Predicted Negative calls (those having a low likelihood of converting (FN/PN - FOR)). (Note that Predicted and Actual Positives (TP/P Recall) and Negatives (FN/N False Negative Rate) are discussed in greater detail below, see Exhibit [E4] for definitions of TP, PP, P, FN, PN, N & etc.) By separating Leads into these two categories, Sales Agents can concentrate their effort and energy on pursuing the Leads who are most likely to buy.

[2] DATA UNDERSTANDING

After downloading the data set, the as-is CSV file was found to be semicolon delimited. Basic data manipulation was performed to get the columns straightened out. Exploratory Data Analysis (EDA) was simple for this data set. It consisted of visual inspection of csv data in excel as well as reading the paper for an overview of prior analysis. Note that the Moro et al paper identifies the top 6 most important input features to their SVM model having the biggest impact on the output target variable. They are:

"Call duration",
"Month of contact",
"Number of previous contacts",
"Days since last contact",
"Last contact result",
"First contact duration".

The data set is held in a CSV file containing a mix of numeric and text data. The rightmost column holds the target variable y which can take on the values 'yes' or 'no' standing for:

"Yes" - "Yes a sale was made as the result of calling this lead." or
"No" - "No sale".

[3] DATA PREPARATION

The data set was loaded into the ipython jupyter notebook via pandas read_csv(). Two batches were created as follows:

BATCH_1 - all features with text data removed, and
BATCH_2 - Moro's top 6 features listed above.

Additional Feature Engineering (FE) was performed in order to prepare the data, including:

FE_1 convert target variable y to numeric from 'yes' and 'no' to 1 and 0.
FE_2 convert month variable from text to numeric (e.g. jan -> 1, feb -> 2)
FE_3 convert poutcome from 'failure', 'nonexistent', 'success' to 1,2,3

The training test split is done in by the usual manner by separating out the four standard sets of training and test features, as well as training and test targets:

[1] X_train - training features
[2] X_test - test features
[3] y_train - training target
[4] y_test - test target

Each model was fit to [1] & [2] training features and targets.

[4] MODELING

The heart of AIML in the CRISP-DM is the [E1d] Modeling phase. See Exhibit [E3] for a high level abstraction/pseudo-code of a Python AIML implementation. This structure was used to implement each of the four classification models to both batches. The four models were introduced earlier, they are:

```
[3a] LGRG - Logistic Regression,
[3b] DT   - Decision Trees,
[3c] KNN  - K-Nearest Neighbors, and
[3d] SVM  - Support Vector Machines.
```

See also the accompanying Jupyter Notebook.

[5] EVALUATION

How effective are the AIML models? Models can be compared to themselves with different input settings as well as one and other by an industry standard set of techniques for evaluating model efficacy. Because these methods can be apply to all models in the same way, models can be cross compared to see which ones are more effective for a given data set. See Exhibit [E4] Definitions for details on the terms used in the Exhibits:

```
[E5] Accuracy
[E6] Precision Recall threshold crossover
[E7] Receiver Operating Characteristic & Area Under the Curve.
[E8] Confusion Matrix
```

Useful shorthand is defined in [E4], and it is good to refer back to this exhibit. For example:

```
TP = True Positive,
P = Positive, and
TP/P = Recall.
```

In Exhibit [E5], the table shows accuracy to be in the high 80s / low 90s across all models, for both training and test sets in both batches: BATCH_1 and BATCH_2. Noting the imbalance between positives and negatives:

TP + FN = 1160 #positive in ys Test
FP + TN = 9137 #negative in ys

The accuracy score being high indicates the total number correctly predicted TP and TN. Because the negatives outnumber the positives ~9:1, Precision vs Recall analysis is performed on for different threshold values. Refer to the table in [E6] as well as the jupyter notebook for Precision vs Recall graphs. The precision recall crossover thresholds are established, and can be seen visually in the graphs. These are marked on the table in [E5] for model intercomparison as well. Although the specific precision recall crossover threshold varies from model to model, most look normal relative to what is expected.

Values for the Receiver Operator Characteristic (ROC) Area Under the Curve (AUC) are given in Exhibit [E7], showing DEC_TRE and KNN to give the best classification for both BATCH_1 and BATCH_2. Decision Trees are particularly effective for BATCH_1 at 99%.

The notebook also has graphical representations of the Confusion Matrix for each model, and a summary table is presented in exhibit [E8]. For BATCH_1, DEC_TRE and KNN are seen to be particularly effective with the highest recall.

[6] DEPLOYMENT

An AIML system such as the one described would be deployed into an organization's production operations to increase Marketing efficiency and productivity. From a high level, many things would have to be taken into consideration in order to adopt the program. From the perspective of IT/Technology, that team would focus on networks, security and computer servers to run the system. The Business Intelligence Analysts would look toward establishing the system, and then making it run better (e.g. increase AIML model performance by tuning and testing hyper parameters). Accounting and Finance would concern themselves with how to plan for and pay for the new program. Last but not least, Marketing would focus on implementation – planning, training and execution.

From the point of view of Marketing Strategy, the impact AIML could have on a business can be seen by looking at two proposed plans PLAN_1 & PLAN_2 (See Exhibit [E8]).

Note that the PBDS is an aggregated summary of internal data which was collected and recorded during the marketing campaign as sales calls were being made. This information was generated and recorded as the campaign progressed. It is aggregated and summarized in the sense that

some information has been lost. For example, the "last call length in minutes" is recorded and saved, however the total length of all calls, as well as the date/time in which they occurred does not appear in PBDS.

Extending the data - It is possible that customers in the PBDS could also be matched to external marketing data as part of the business process. For the sake of this discussion, consider these to be Third Party Data Sets that could be purchased from some external third party data vendor. These "Leads Lists" or "Leads" could be joined to the PBDS for the purposes of training models, and then additional Leads Lists could be purchased to forecast future sales. Their resulting outcomes could be recorded as the PBDS continues to grow, and the sales conversion result continued to be recorded in the target variable y. The forecasts would be comprised of a set of predicted positives, and these are the Leads who are most likely to convert. Continuing in this way, two business plans emerge (see Exhibit [E8]) are similar and based on the following sequence:

[1] Buy new Leads Lists,
[2] Model and Predict
[3] call Predicted Positives (PP)
[4] call Predicted Negatives (PN)

The only difference between PLAN_1 and PLAN_2 is the inclusion of step [4]. Whereas PLAN_1 is the first three steps [1] - [3], the second plan PLAN_2 would be all four steps [1] - [4]. This simple business process can be used to achieve strategic sales objectives as new Leads continue to be purchased, modeled and converted to new customers. Consider the following definitions:

TAM Total Addressable Market
TOL Total Obtainable Leads Lists
DNNC Desired Number of New Customers (Sales Target)

as well as:

PPY Predicted Positive Yield (aka - Precision or Positive Predicted Value (PPV))
PNY Predicted Negative Yield (aka - False Omission Rate (FOR))

where PPY is the number of TP relative to PP. Also consider the equation:

$$PPY * TOL > DNNC$$

which indicates that the Sales Targets in DNNC can be reached by calling all PP from the Leads Lists, when the size of those lists and the PPY in combination exceed the target number. Consider the following scenarios:

If buying Leads is inexpensive relative to calls, then PLAN_1 is more cost effective.
If buying Leads is expensive relative to calls, then PLAN_2 is more cost effective.

Other factors will also play into the situation and will need to be considered when assessing the overall cost efficiency of the program. These and other business considerations would have to be discovered as a campaign proceeds and the specific realities of that campaign come to be known. That said, the proposals above form a solid foundation for implementing the AIML methods, meeting sales objectives with greater efficiency and responding observed and decided as the campaign develops.

CONCLUSION

In conclusion, the marketing campaign as well as the data record in PBDS is an excellent start to customer acquisition. The PDBS can be used with AIML methods for business intelligence analytics that will make marketing more efficient, effective and productive.

The key observation and takeaways are the following. Buying Leads lists and running AIML models to predict sales can increase marketing effectiveness. Calling the Predicted Positive PP Leads can increase call throughput to as high as one sale for every 2 calls. Calling the Predicted Negative PN Leads could have call throughput as low as 1 sale for every 20 calls. Thus AIML conversion throughput is 10X higher for PPs than for PNs.

The AIML methods as a whole are used to predict sales calls that will result in a positive conversion. Combining this with the ability to purchase customer profile data are key concepts and insights. Together these proposals represent an easy, straight forward way to understand, apply and leverage the AIML tools as a means to higher profitability.

EXHIBITS

[E1] EXHIBIT_1

Listing [L1] an outine of Business and AIML concepts.

[1] CRISP-DM - Data Mining Six Phase Cycle

- [1a] Business Understanding
- [1b] Data Understanding
- [1c] Data Preparation
- [1d] Modeling
- [1e] Evaluation
- [1f] Deployment

[2] EQ_1 Accounting Equation

- [2a] Assets = Liabilities + Equity
- [2b] A = L + E

[3] EQ_2 Gross Profit, Finance Equation

- [3a] Gross Profit = Revenue - Cost
- [3b] Gross Profit = Revenue - (Marketing Cost + All Other Costs)
- [3c] GP = R - (MC + AOC)

[4] Customer Lifetime Value

- [4a] Acquisition,
- [4b] Monetization, and
- [4c] Retention.

[5] Marketing Four P's

- [5a] Product
- [5b] Placement
- [5c] Pricing
- [5d] Promotion

[6] Purchase Funnel

- [6a] Awareness
- [6b] Interest
- [6c] Desire
- [6d] Action

[7] Operations Management, Theory and Principles

[7a] Process Efficiency
[7b] Value Stream Mapping
[7c] Quality (increase), 7 tools
[7d] Speed (increase)
[7e] Cost (decrease)
[7f] Process Mining (i.e. workflow analysis)

[8] AIML Artificial Intelligence & Machine Learning

[8a] PBDS Portuguese Bank Data Set
[8b] Feature Engineering - two batches.
[8c] Test Training Split
[8d] AIML Model - fit training data
[8e] AIML Model - predict training data
[8f] AIML Model - predict test data
[8g] AIML Model - compare Test prediction vs Test Actual
[8h] AIML Model - assess efficacy via ROC_AUC, confusion matrix
[8i] Applications - institutionalize the AIML model into the operating plan and daily work activities according to corporate strategy.

[9] Production Operations - Kaizen, Continuous Improvement

[9a] TQM/TQI Total Quality Management, Total Quality Improvement,
[9b] TQM Phases 1, 2, 3 & beyond,
[9c1] Phase 1 pre AIML - establish baseline,
[9c2] Phase 2 initial AIML - early understandings & application,
[9c3] Phase 3 evaluate [8d] Phase 2 and plan improvements,
[9cR] Repeat phases 2 & 3, introduce new change, see the effect, plan and implement next change.
[9d] AIML model findings prototype/pilot,
[9e] AIML model build for production,
[9f] AIML model deployment,
[9g] SOP checklists, training, maintenance,
[9h] Phase 3 production - observe & evaluate,
[9i] Beyond - Maintain & Sustain the Continuous Improvement,
[9j] Study literature (e.g. LSS, PDCA, OODA, TOC),
[9k] SOP - standard operating procedures (i.e. checklists)
[9l] process improvement LSS, PDCA, OODA, TOC
[9m] Continuous Improvement - apply and improve model findings
[9n] TODO think about how to incorporate LSS OODA, PDMA
[9o] TODO LSS Lean Six Sigma & Theory of Constraints. LSTC

[E2] EXHIBIT_2 Feature Engineering

The Feature Engineering consisted of producing two batches:

[1] BATCH_1 - remove strings etc , and
[2] BATCH_2 - highest impact features (see paper page 5 & 6).

The full list of columns from the data set. Note that 0-19 are features and labeled 'F' and the column containing the target variable is 'T'. For features, their inclusion in a batch is marked by an 'x', and their exclusion marked by a dot '.'.

		COLUMNS	BATCH_1	BATCH_2
F	0	age	x	.
F	1	job	.	.
F	2	marital	.	.
F	3	education	.	.
F	4	default	.	.
F	5	housing	.	.
F	6	loan	.	.
F	7	contact	.	.
F	8	month	.	x
F	9	day_of_week	.	.
F	10	duration	x	x
F	11	campaign	x	.
F	12	pdays	x	x
F	13	previous	x	x
F	14	poutcome	.	x
F	15	emp.var.rate	x	.
F	16	cons.price.idx	x	.
F	17	cons.conf.idx	x	.
F	18	euribor3m	x	.
F	19	nr.employed	x	.
T	20	y	T	T

[E3] EXHIBIT_3 Abstraction of Machine Learning process as represented in code.

The basic sequence of steps in a simple Machine Learning python program:

```
    df = load(data)
    X,y = FeatEng(df)
    Xr, Xs, yr, ys = test_train_split(X,y)
    mod = Model()
    mod.fit(Xr,yr)
    yps = mod.predict(ys)
    acc = accuracy_score(yps,ys)
    prc = precision_score(yps,ys)
    rcl = recall_score(yps,ys)

where:
    df = DataFrame
    X is a feature matrix, and
    y is an outcome vector.

    X  = Xr  + Xs
    y  = yr  + ys
    yp = ypr + yps

    Xr   X features training
    Xs   X features test
    yr   y actual training
    ys   y actual test
    ypr  y predeicted training
    yps  y predeicted test
    acc  accuracy measure
    prc  precision measure
    rcl  recall measure
```

[E4] EXHIBIT_4 Definitions

Definitions for Predicted and Actual Positive/Negative target variable outcomes. The following defines abbreviations and nomenclature for industry-standard naming conventions used for statistical ratios with analytic meaning. See also article references for further details.

```
P = Positive
N = Negative
π = Predicted
α = Actual
1 = 'Yes'  Lead converts to a sale
0 = 'No'   No sale

π α
1 1  TP = True  Positive
1 0  FP = False Positive
0 1  FN = False Negative
0 0  TN = True  Negative

π = 1  PP Predicted Positive TP+FP
π = 0  PN Predicted Negative TN+FN
α = 1  AP Actual Positive
α = 0  AN Actual Negative

TP/P        Recall - TP Rate, Sensitivity
FP/N        Specificity - FP Rate
TP+TN/P+N   Accuracy
TP/PP       Precision
FN/PN       FOR - False Omission Rate
```

Tom Fawcett _An Introduction to ROC analysis_
https://www.sciencedirect.com/science/article/abs/pii/S016786550500303X
https://en.wikipedia.org/wiki/Confusion_matrix
https://en.wikipedia.org/wiki/Sensitivity_and_specificity

[E5] EXHIBIT_5 Accuracy. Table showing model accuracy across both batches for both training and test sets.

RPT_ACCURACY	BATCH_1	BATCH_2
svm_accu_train	0.898	0.905
svm_accu_test	0.896	0.902
knn_accu_train	0.933	0.926
knn_accu_test	0.904	0.896
lgrg_accu_train	0.910	0.905
lgrg_accu_test	0.905	0.904
dec_tre_accu_train	0.999	0.961
dec_tre_accu_test	0.890	0.890

[E6] EXHIBIT_6 Precision Recall threshold crossover for BATCH_2. See also the Precision Recall graphs in the jupyter notebook. These values were collected and compiled from that source.

	SVM		KNN		LGRG		DEC_TRE	
THRS	PREC	RCLL	PREC	RCLL	PREC	RCLL	PREC	RCLL
0.0	0.11	1.00	0.11	1.00	0.11	1.00	0.11	1.00
0.1	0.58	*0.44<<	0.32	0.80	0.36	0.73	0.32	0.50
0.2	0.63	0.35	0.32	0.80	0.51	0.55	0.38	0.48
0.3	0.64	0.30	0.46	0.59	0.57	*0.46<<	0.43	0.46
0.4	0.66	0.28	0.46	0.59	0.60	0.39	0.46	*0.43<<
0.5	0.67	0.25	0.55	*0.39<<	0.64	0.34	0.47	0.43
0.6	0.71	0.22	0.63	0.23	0.65	0.25	0.52	0.37
0.7	0.69	0.18	0.63	0.23	0.67	0.16	0.53	0.37
0.8	0.71	0.13	0.63	0.23	0.69	0.10	0.53	0.37
0.9	0.79	0.05	0.75	0.10	0.63	0.05	0.53	0.37

where:

```
*0.00<< First threshold after precision and recall crossover.
THRS    Threshold
PREC    Precision
RECL    Recall
```

[E7] EXHIBIT_7 Receiver Operating Characteristic (ROC) Area Under the Curve (AUC). Table showing ROC AUC for both batches. Based on ROC AUC the Decision Trees came in as the best most effective models, and the KNN models were next best.

ROC AUC	BATCH_1	BATCH_2
svm_roc_auc	0.603	0.634
knn_roc_auc	0.799	0.759
lgrg_roc_auc	0.689	0.655
dec_tre_roc_auc	0.999	0.836

[E8] EXHIBIT_8 Confusion Matrix. Table for Confusion Matrix values for Test set predicted by different models on different batches. The [m4.1] Batch_1 Decision Tree Model clearly has the highest model efficacy from the perspective of TP/PP Precision at 52.4%. And the least effective is [m1.1] Batch_1 at 22.6%.

	MODEL	BATCH	TP	FP	FN	TN	PP	PN	PREC TP/PP	FOR FN/PN
[m1.1]	svm	1	251	153	909	8984	1160	9137	21.6%	9.9%
[m2.1]	knn	1	537	365	623	8772	1160	9137	46.3%	6.8%
[m3.1]	lgrg	1	431	248	729	8889	1160	9137	37.2%	7.9%
[m4.1]	dec_tre	1	608	573	552	8564	1160	9137	*52.4%<<	6.0%
[m5.2]	svm	2	329	175	831	8962	1160	9137	28.4%	9.0%
[m6.2]	knn	2	463	372	697	8765	1160	9137	39.9%	7.6%
[m7.2]	lgrg	2	399	218	761	8919	1160	9137	34.4%	8.3%
[m8.2]	dec_tre	2	439	407	721	8730	1160	9137	37.8%	7.8%

FOR = FALSE_OMISSION_RATE
PREC = PRECISION

[E9] EXHIBIT_9 Business Plan - Applying the findings.

Because this exercise occurs is an economic and business context, it is oriented toward decision making to get the most benefit as quickly as possible at the least cost. Because there are so few PP (Predicted Positives) relative to PN (Predicted Negatives), and the likelihood of a sale is so much higher for the PPs, the guidance is to start by calling all of the PPs first. This is the best way to gain the most sales in the least amount of time. Calling the PPs and only the PPs will have the highest sale to call throughput. Said another way – the idea is to get a set of Leads, call all of the Predicted Positives, and then the sales yield would be the Precision. The next step would be to obtain another list and repeat by once again calling all of the Predicted Positives. If lists are expensive, then calling all of the Predicted Negatives would be advisable, however the average number of calls to achieve a sale would be slower overall. Marketing Managers would have to weigh the tradeoffs of obtaining datasets and making calls with their associated costs. To that end, harvesting the AIML results as given in [E8, m4.1] above:

```
TP/PP - Precision       - yield 52.2% - 1:2  calls convert to sale.
FN/PN - False Omissions – yield  6.0% - 1:20 calls convert to sale.
```

Buying lists and making calls can be can be expensive or inexpensive, both in absolute terms, as well as relative to each other, also in terms of the relative number of converted customers per data set, or converted customers per 1000 calls. Thus, depending on conditions, one of the two different plans can be employed. The plans are:

```
PLAN_1 - Call Positives, discard negatives, get another data set.

[p1a] given a new data set
[p1b] Run all models
[p1c] predict positives and negatives
[p1d] call all positives PPs are new subscribers (ignore PNs)
[p1e] discover precision: TP / PP, calculate yield
[p1f] obtain new data set, start again at [p1a].


PLAN_2 - Call Positives, call negatives, then get another data set.

[p2a] given a new data set
[p2b] Run all models
[p2c] predict positives and negatives
[p2d] call all positives PPs are new subscribers
[p2e] call all negatives PNs are new subscribers
[p1f] discover precision: TP / PP, calculate yield
[p2g] discover false omission: FN / PN, calculate yield
[p2h] obtain new data set, start again at [p1a].
```

When datasets are inexpensive and calls are expensive do plan [p1]. When datasets are expensive and calls are inexpensive do plan [p2]. When both datasets and calls are either expensive or inexpensive do plan [p1], unless there is a good reason to go with plan [p2].

It will be fundamental for the success of the program to determine the cost of exhaustively calling PN relative to cost of buying new Leads. Depending on affordability, effort and availability, purchasing new Leads Lists could be less expensive than calling all predicted PNs. The question is - which course of action will have the highest throughput rate leading to the most sales at the least cost in the least amount of time. Understanding these dynamics, and answering these questions would be the concerns of [E1.7] Operations Management in the context of future marketing campaigns. For the purpose of business application, there is no reason not to do the feature engineering for both batch_1 and batch_2 and then run all four models across both batches thereby harvesting all of the Predicted Positives regardless of the model from which those PPs were derived. By taking the union of all Predicted Positives found in both batches across all models, this total calls list represents the most likely and least expensive path to gaining as many sales as possible in the shortest amount of time.