Mining for Gold: How to Predict Service Contract Performance with Optimal Accuracy based on Ordinal Risk Assessment Data

Sinem Güven[1], Mathias Steiner[1], Tsuyoshi Idé [1], Sergey Makogon[2], Alejandro Venegas[3]

[1] IBM T. J. Watson Research Center, Yorktown Heights, NY, USA
[2] IBM Global Technology Services, Salt Lake City, UT, USA
[3] IBM Global Technology Services, Santiago de Chile, CHILE

{sguven, msteine, tide, smakogon}@us.ibm.com, avenegas@cl.ibm.com

Abstract

Proactive management of service contract risks ahead of contract signing is becoming increasingly important for IT service providers due to the cost pressure associated with IT outsourcing. Within an end-to-end risk management process, various risk assessments are performed at multiple stages before a service contract is signed. Based on the risk assessment data, service providers seek to have predictive models that indicate risks of future service contracts. Considering the wide range of risk assessments, the variable frequency in which they are conducted, their sequential nature and the prevalent data scale, naïve statistical modeling approaches, such as linear regression, are not readily applicable to such data sets. It is, therefore, necessary to identify a new methodology for predicting service contract risks based on ordinal risk assessment data.

In this paper, we describe an analytical methodology that enables optimal risk prediction for service contracts, along with the lessons learned from implementation within an enterprise-level risk management ecosystem. Such real-world insights can provide guidance to data scientists and researchers both in the service delivery domain as well as other domains with similar data characteristics.

1. Introduction

The growing trend of Big Data [1] enables organizations to drive innovation through advanced predictive analytics that provide new and faster insights into their customers' needs. According to Gartner [2], by 2016, 70% of the most profitable companies will manage their businesses using real-time predictive analytics. Incidentally, IT service providers are relying more and more on predictive analytics for advanced risk management [3]. Such analytics enable service providers to predict risks ahead of time and proactively manage them to eliminate or minimize their impact.

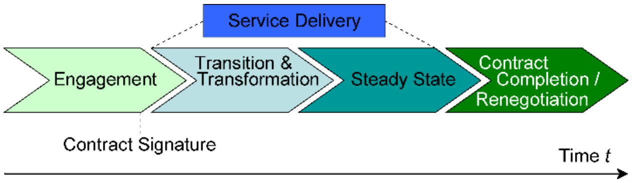

Figure 1. IT outsourcing service lifecycle

This paper describes our experience and methodology for improving the accuracy of contract risk prediction through optimization of the data selection process. Our scope is managing the risk of *IT outsourcing contracts* (or *service contracts*), however, the methodology we present can be used in other domains with similar data characteristics (see Section 2).

Figure 1 shows the typical lifecycle of a service contract. Predictive analytics can help in the *Engagement* (or *pre-contract*) phase to make informed decisions about whether to sign a risky contract as well as how much contingency should be included in the contract price [4]. In *Transition and Transformation (T&T)*, where the IT service provider transforms the client's infrastructure and operations into a format they can effectively manage, predictive analytics can provide insights into operational risks based on historical data to help proactively mitigate those risks. In *Steady State*, where the outsourcing service reaches maturity but there is less tolerance for failure, predictive analytics can be used to detect and prevent system failures [5]. Predictive analytics is, thus, integrated into various steps within the end-to-end risk management process.

Throughout the service contract lifecycle, risk management insights are typically collected through surveying risk managers or quality assurance experts [6]. Such risk assessment data, which mainly comprises ranked score values, is a valuable source for predictive analytics as it already captures the *status quo* of the contract at hand. For service contracts, risk assessment surveys are typically conducted at variable time intervals depending on the complexity of the project. The more complex the project is the earlier the risk management is involved, and the more often the risk assessments are conducted. There may be several different types of risk assessment surveys some of which include but are not limited to: a) Technical Assessment; b) Client Assessment; c) Solution Assessment. Throughout the lifecycle of a service contract, several risk managers and independent quality assurance experts perform these surveys to ensure that input is collected from all perspectives. Thus, the same survey is repeated several times across different time ranges.

During the *Service Delivery* phase, which contains both *T&T* and *Steady State* (Figure 1), service providers track the performance of outsourcing contracts through different *Key Performance Indicators* (KPIs). Just like the risk assessment surveys, KPIs are also collected at variable time intervals depending on the complexity and the health of the contract. The more troubled the contract is, the more attention it will need and the more often the KPIs will be measured and updated.

Within the Service Delivery domain, one of the main applications of analytics is to predict one or more of such KPIs in the Engagement phase in order to reveal contractual issues as early as possible. When building a risk model for predicting contract performance, even if we focus on a specific risk assessment as *input* and a specific KPI as a *target*, there is still a wide range of inputs and targets to choose from with variable time delays in between. It is, however, unclear which data selection criteria should be applied to narrow down the scope, or how data selection affects prediction accuracy. Another important issue with the IT outsourcing data is that, due to its inherent characteristics (described in Sections 2 in detail), naïve statistical modeling approaches, such as linear regression, are not readily applicable.

This paper addresses the above issues by proposing a novel methodology for building optimal predictive models from complex IT outsourcing data sets. We first analyze how the training data set selection with respect to the time delay between risk assessments and contract performance measurements (KPIs) affects the accuracy of contract risk predictions. We then show how accuracy of prediction models can be optimized through insights gained from such analysis.

In the following Section, we discuss the characteristics and the inherent complexity of the IT contract risk data.

2. Data Characteristics and Complexity

We use the term *Contract Risk Assessment* (CRA) to refer to the service contract risk assessment surveys. The term *Contract Performance Measure* (CPM) comprises a single KPI, or several KPIs merged together through a business logic to track contract performance. As described in Section 1, CRA and CPM data are collected across different stages of the contract lifecycle at varying frequencies and time intervals depending on the complexity of the contract. Figure 2 shows the time distributions of the contract risk assessments ($CRA_{1,....,n}$) and the contract performance measures ($CPM_{1,...,m}$).

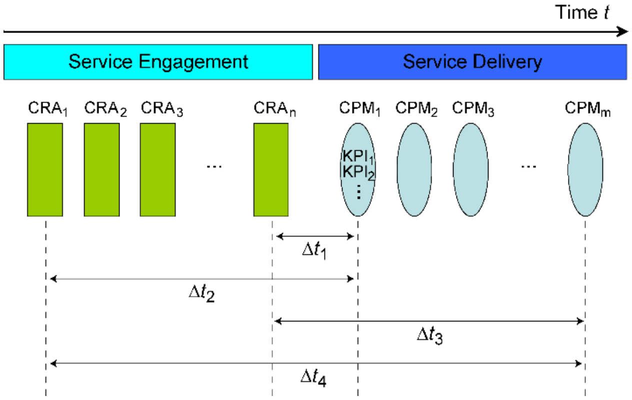

Figure 2. Timeline of contract risk assessments (CRA) & contract performance measures (CPM)

2.1. Data Types: Risk Model Input (CRA) and Target (CPM)

CRA data is generated through surveys, which vary from 20-200 questions. Each survey question typically has a variety of categorical answers to choose from, which range from high to low, or vice versa. For each such survey, there is an underlying algorithm, which calculates a final risk assessment score based on question answers.

CPMs, on the other hand, can be in the form of a *survey* (in which case an underlying algorithm calculates a CPM score) or an *actual measurement* (such as the Gross Profit of the contract for that month). As mentioned earlier, CPMs may represent one or several KPIs.

2.2. Data Characteristics

2.2.1. Variable Time Delay

CRA and CPM data may not necessarily come from periodic assessments, but rather varying time frames (as they are conducted on an as-needed basis). This means that there is a variable time delay between CRA and CPM data rendering some data points potentially irrelevant due to major time lag.

2.2.2. Incomplete Data

CRA and CPM data may contain blanks as not all assessment questions and performance measures are mandatory.

2.2.3. Evolving Data

The needs of the business and the associated risks change over time, thus requiring changes in the risk assessment questions and performance measures. For CRA, this results in surveys with modified or new questions. For CPM, the definition of the performance measures may change or new measures may be added.

The unique combination of data characteristics described above render predictive modeling for IT outsourcing a non-trivial task. The next few Sections describe our experience and methodology on how to build an optimal predictive model based on such data sets.

3. Business-Rule Driven Data Selection and Predictive Model Conception

As described in Section 1, our ultimate goal is to predict KPIs reliably using CRA data at Engagement time. For the purposes of this paper, we focus on financial profitability of a contract, and choose to predict the *Gross Profit Variance* KPI denoted by K(ΔGP). This numeric KPI is defined as the projected gross profit minus the actual gross profit.

The first step to building a predictive model is to perform training data selection from our historical data set. As described in Section 2, even if we limit ourselves to one type of CRA as our *input*, and the K(ΔGP) as our *target*, we still have a wide range of input and target variables to choose from as CRAs and KPIs are measured several times across the service contract lifecycle. To better illustrate the complexity of our data selection problem, imagine hundreds of historical contracts, each of which have several iterations of the chosen CRA, and similarly several measurements of the chosen target KPI, K(ΔGP). This means that, for each historical contract, our training data should include the one CRA and the one K(ΔGP) that best represents that historical contract's risks and observed Gross Profit Variance respectively. Populating the training data set with the right CRA and K(ΔGP) instances for hundreds of historical contracts is, hence, a major challenge.

Our initial attempt to select the training data set that best represents the historical contracts relies on a business driven approach. We decided to use the *last CRA* of each historical contract as input, as it is closest to the contract signature, and thus, reflects the best risk information known about the contract. As for the target variable, we decided to choose the *K(ΔGP) closest to the end of T&T*, as it best reflects T&T performance. Our predictive model would, thus, predict what the financial profitability of a contract would be at the end of T&T by taking in, as input, the last CRA conducted just before contract signing.

3.1. Accuracy Metrics

Based on business needs, we defined 3 metrics to assess the accuracy of our model:

- *Directional accuracy*: how accurately the model predicts whether a new opportunity will become profitable, or not
- *Non-profitable contract prediction accuracy (NPCP):* how accurately the model predicts the opportunities that will become non-profitable
- *Profitable contract prediction accuracy (PCP)*: how accurately the model predicts the opportunities that will become profitable

Although the main intent of such predictive models is to achieve high classification accuracy for non-profitable contracts, the accuracy of the profitability prediction is just as important. Without a high PCP accuracy, false negative predictions would lead to unnecessary risk mitigation activities in healthy contracts.

3.2. Correlation between Input and Target Data

After data selection has been performed, the next step is to investigate which input variables (CRA questions) have significance for predicting the target variable K(ΔGP). To achieve this, we calculate Pearson's Correlation between various sets of CRA questions and the K(ΔGP)s of historical contracts and pick the combinations with the highest occurring correlations.

As expected, correlations between input and target variables are typically strong when the CRA is specifically designed to determine the target KPI. Although there are several different questions that aim to capture financial performance within the CRAs of our data set, none of the CRAs was designed for determining contract profitability specifically. As a result, we observe only a weak correlation ($r <= 0.25$) between the input and target variables. Given the extensive set of questions the practitioners typically answer during contract risk assessment, it is unpractical to add dedicated CRAs for each KPI of interest. Instead, our objective is to predict KPIs based on the *available* data with optimal accuracy.

3.3. Predictive Model Conception

As discussed in previous Sections, IT outsourcing data has inherently unique characteristics, which render naïve modeling techniques, such as linear regression, not readily applicable. The *sequential* nature of the survey data precludes the assumption of statistical independence between observations. Moreover, the *ordinal* scale level of survey data means that statistical models that require interval or ratio scale levels are not suitable. Furthermore, even if ordinal regression (an extension of the general linear model to ordinal categorical data) were to yield a model with reasonable accuracy, it is often difficult to straightforwardly interpret the meaning of individual regression coefficients. For example, a risk model might entail a negative regression coefficient associated with a particular CRA risk question. Reduction of the risk score associated with that CRA question might lead to an increased Gross Profit Variance (less profitability), which is a counter-intuitive risk prediction from the practitioners' perspective. In addition, most naïve modeling techniques do not perform well on data sets with *blank entries*, a major characteristic of the IT outsourcing data sets. Finally, it is difficult for such models to automatically re-train or evolve with the *changing data sets*.

3.3.1. k-Nearest Neighbor (KNN)

Considering the above limitations, we decided to use the *k*-Nearest Neighbor (KNN) [7] approach to predict K(ΔGP). Unlike most other modeling techniques, such as linear regression, KNN does not rely on a specific parametric model. Instead, it simply uses *k* historical contracts that are most similar to the new opportunity to predict the K(ΔGP) for that opportunity. Since each prediction is represented by the most similar historical contracts, KNN allows highly interpretable results. Also, thanks to the nonparametric nature, KNN can handle complex nonlinear relationships between input and target variables. More importantly, KNN has the flexibility to be tailored to business requirements by customizing the notion of similarity.

3.3.2. Question Importance

When calculating contract similarity, our model uses the correlation between input and target variables (Section 3.2) as *weights*, which are indicators of the importance of CRA questions in determining a contract's K(ΔGP).

3.3.3. Full Parameterization

The predictive model is fully parameterized to enable identification of various optimal thresholds that maximize the model's performance:

- *Question_Importance* threshold is used to ensure that only the most relevant CRA questions are ultimately used in determining contract similarity.
- *Contract_Similarity* threshold is used to determine the minimum degree of similarity a historical contract should have to the new opportunity before it can be included in the K(ΔGP) prediction.
- *Outliers* parameter is a Boolean that determines whether outliers beyond a defined observed K(ΔGP) range should be included or excluded from the calculations considering the vast range of observed K(ΔGP)s in historical data.

3.3.4. Calculating Gross Profit Variance: K(ΔGP)

Once contracts similar to the new opportunity are identified, we take a weighted average of their observed K(ΔGP)s by taking their degree of similarity into consideration to determine the final K(ΔGP) prediction for the new opportunity, as shown in Equation 1.

$$
\boxed{\begin{array}{c} K(\Delta GP) = \sum (\Delta GP\_Actual_i * contract\_similarity_i) / total\_similarity \\ i=1 \text{ to } N \text{ where } N = \# \text{ of similar contracts} \end{array}}
$$

Equation 1. Weighted average based on contract similarity

3.4 Initial Predictive Modeling Results

We trained and tested our KNN model on the selected data set (see Section 3). Testing was performed through Leave-One-Out Cross Validation (LOO-CV) with over 900 different threshold configurations to maximize model accuracy. The results are shown in Table 1. While the model yields a reasonable accuracy for *Non-Profitable Contract Prediction (NPCP)* at 71%, it fails on *Directional* and *Profitable Contract Prediction (PCP)* metrics.

Table 1. Initial model accuracy based on business driven data selection

Directional	NPCP	PCP
59%	71%	52%

4. Data-driven, Optimized Data Set Selection and Predictive Model Enhancements

In order to overcome the limitations of the business-rule driven data selection approach discussed above, we decided to select the training data set through a data-driven methodology based on machine learning techniques [14].

4.1. The Methodology

Our optimal data selection methodology entails the following steps:

1) Determine if *time delay* has any significance in selecting training data, given the wide range of input and target variables with varying time frames. For example, if a given historical contract has the same CRA repeated several times, understand if using the first one vs. the last one has any effect on the accuracy of models trained with such CRAs.

2) If time delay does have significance, select the *optimal time window* in the data set. Once the optimal data set is selected, train the predictive model using this data set to maximize prediction accuracy.

At a high level, the problem of selecting optimal data set resembles the well-known research areas of *Feature Selection* and *Sample Selection*. Feature selection [8, 9] refers to algorithms that select a subset of the input data features that performs best under a certain classification system. Our approach is different from feature selection as we are selecting the optimal time window (based on the entire available data set) by monitoring and maximizing the prediction accuracy of the risk models irrespective of the number of features.

Sample selection [10], on the other hand, is focused on how to achieve a good accuracy for a predictive model with a reasonable number of sample points. The accuracy of a predictive model is to a large extent determined by the modeling technique used, but the sample selection often has a direct influence on the model performance [11]. Unlike common sample selection techniques, the main goal of our work is not to optimize the number of sample points, but the time distribution of the modeling data set to achieve maximum prediction accuracy in the resulting risk models, independent of the modeling algorithm used.

4.2. Data-driven Data Set Selection

4.2.1. Data Preparation

In order to test the significance of *time delay* in determining an optimal training data set, we have chosen the following data clean-up criteria:

- Exclude incomplete CRAs and CPMs: we do not perform any data filling [12] so as not to introduce any bias to the data.
- Exclude unique survey questions (that are not part of all CRAs or CPMs - if they are in the form of a survey): we do not perform any question mapping so as not to introduce any bias to the data.
- Exclude temporal inconsistencies: we calculate the time difference between CRAs and CPMs and exclude those CRA-CPM combinations with a negative time delay (i.e. CPM performed before CRA)

Based on the above criteria, the data clean-up is performed via Java and CLEM (SPSS Modeler). We then identified four data sets that represent different time delays Δt_i between CRAs and CPMs, as shown in Figure 2. Since risk assessment results and service contract status are subject to change over time, it is reasonable to assume that the accuracy of predictive models trained on the data will critically depend on the time delay between them. Nevertheless, other data selection criteria such as the risk assessment outcome or the performance measurement result, e.g. best case versus worst case, could also be considered.

The data set characterized by Δt_1 connects for each service contract the last risk assessment performed (CRA_n) with the first performance measure conducted (CPM_1). Similarly, the data set characterized by Δt_2 connects the first risk assessment (CRA_1) with the first performance measure (CPM_1) while Δt_3 represents the data set that associates the last risk assessment (CRA_n) with the last performance measure (CPM_m). Finally, Δt_4 characterizes the data set that correlates the first risk assessment (CRA_1) with the last performance measure (CPM_m).

4.2.2. Approach and Implementation

Our data-driven model uses 24 question answers of the chosen CRA (see Section 3.2) as predictor input. The dichotomous model target K(ΔGP): *Gross Profit Variance* is derived from the actual performance measures as described in Section 2.

To generate predictive risk models based on machine-learning algorithms [13], the data is partitioned such that 70% of data is used for model training and 30% of the data is used for model testing. For comparison, we alternatively apply C5.0 and Binomial Logistic Regression algorithms (through IBM SPSS Modeler) while maintaining the initial modeling conditions. Based on the input data set (CRA), the models classify the target output (K(ΔGP)) and the overall modeling accuracy is compared based on the number of correctly classified service contracts for both training and testing data sets.

The frequency distributions of the selected data sets (each contains N_1=164 projects) are plotted in Figure 3 as a function of the time delay $\Delta t_{\text{CRA-CPM}}$ between risk assessments (CRA) and contract performance measures (CPM).

As expected from looking at Figure 2, the frequency distributions of the different data sets, shown in Figure 3, display different center values. The data set represented by Δt_1 displays relatively short delay times with a distribution median of 4 months. The median of the Δt_2-distribution is twice as high, about 8 months. The medians of the distribution characterized by Δt_3 and Δt_4 are 22 months and 27 months, respectively. Along with the increase of median delay times, we observe a broadening of the frequency distributions.

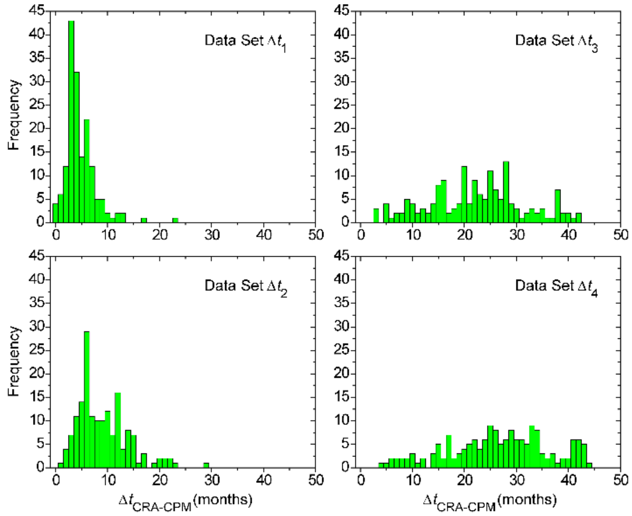

Figure 3. Distributions of time delays between CRA and CPM

4.2.3. Data Selection Criterion: Time Delays between Input Data (CRA) and Target Data (CPM)

In this Section, we investigate how the modeling accuracy is affected by time delays for the modeling target K(ΔGP), which constitutes the relevant financial performance metrics at the time of the contract performance measurement.

The modeling results for K(ΔGP) obtained with the different data sets are listed in Table 2. The classification accuracies for the training data sets vary by less than 10% (with the exception of the model trained with C5.0 on data set Δt_4). More importantly, the prediction accuracies increase as function of the median time delay Δt_i that characterizes the data sets.

Table 2. Modeling results for K(ΔGP) obtained with data sets characterized by different time delays

Data Set	Mean/Median Time Delay (months)	C5.0 Classifier Training/ Testing Accuracy (%)	Logistic Regression Training/ Testing Accuracy (%)
Δt_1	6.5 / 5.2	89 / 49	79 / 59
Δt_2	10.8 / 9.9	85 / 59	80 / 62
Δt_3	22.1 / 21.9	81 / 67	79 / 69
Δt_4	26.3 / 26.0	67 / 74	75 / 69

The classification results obtained with the testing data sets are plotted in Figure 4. The dashed line serves as a guide to the eye; the different data sets are marked by the respective time delays Δt_i (comp. Figures 2, 3 and Table 2).

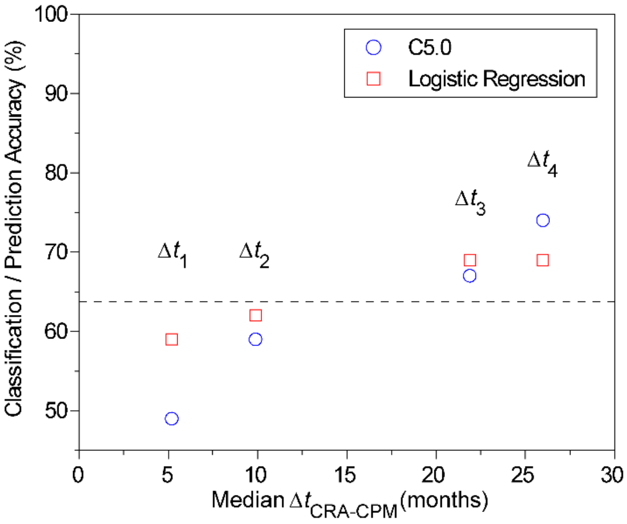

Figure 4. Accuracy of predictive models for the K(ΔGP) (testing data sets, see Table 2)

We observe that the accuracy of models that predict K(ΔGP) is higher for larger median time delays (see Figure 4). The finding that the prediction accuracy for K(ΔGP) increases as function of the time delay between contract risk assessments and contract performance measures can be rationalized by taking into consideration the strong ΔGP-fluctuations that typically occur at early stages of service delivery.

This study also confirms that the original attempt at selecting the training data set based on business insights alone (see Section 3) was not optimal as selecting the last CRA naturally minimizes the time delays between the input and target variables. However, the data-driven selection method discussed here is used only to detect the dependence of classification accuracy on time delay between CRA and CPM. In order to find the optimal time window for training the risk model, further data analysis is needed.

4.3. The Optimal Data Set for Maximizing the Prediction Accuracy of the Predictive Risk Model

With the significance of time delays confirmed, the next step is determining the optimal time window within the data set that would maximize the accuracy of predictive models. First, based on business rules, we select 3 time windows each for Engagement and Delivery timeframes. Engagement windows are labelled E_*N* and Delivery windows are labelled D_*M* respectively, where *N* and *M* represent one of {1, 2, 3}, as shown in Figure 5.

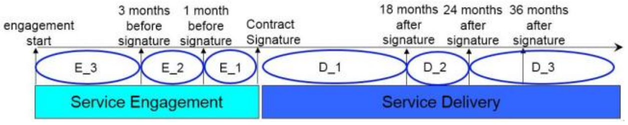

Figure 5. Engagement and Delivery time windows in the SO lifecycle

We then generate training samples by taking a combination of {Engagement X Delivery} windows. This yields 9 *Time Window Combinations (TWC)* and associated data sets to train the model with; matching N^{th} Engagement window with the M^{th} Delivery window, and so on.

Once these data sets are generated, we determine the optimal {Engagement X Delivery} TWC by evaluating the informativeness of each TWC. Our approach is based on the notion of statistical two-sample tests. For each of the training data sets belonging to the 9 TWCs, we first separate the historical contracts into two groups depending on the directionality (positive or negative) of their Gross Profit Variance. We then evaluate the difference between probability distributions of the historical contracts' CRA questions. To quantitatively measure the distributional distance, we use the single-variable Kolmogorov–Smirnov (KS) statistics [14] averaged over the CRA questions. The bigger the averaged KS statistic is, the more informative the TWC is. If there is no significant difference between the positive and negative Gross Profit Variance groups, we conclude the TWC is not informative.

The results of KS statistics revealed that the most optimal TWC is {E_2 X D_3} (Figure 5). This result can be rationalized if we consider that E_3 window is too early in the contract's lifecycle to have all the relevant risks captured through CRAs. E_1 is likely the least risky CRA given that at least some, if not most, risks are identified and mitigated by this stage, leaving E_2 as the time window that represents most relevant risks through its CRAs. Similarly for Delivery data, the optimality of D_3 could be rationalized by considering the early fluctuations of Gross Profit Variance during D_1 and D_2 time windows.

Finally, it is noteworthy to mention the consistency between the KS statistics (Section 4.3) and the time delay study (Section 4.2) results. We observe that the optimal {E_2 X D_3} TWC belongs to the Δt_4 window, which was determined to be the most optimal for our data set.

4.3.1. Weights

Aside from determining the optimal time window, another important consideration for our data set was the low correlation between the input and target variables, as discussed in Section 3.2. Our model uses the correlation coefficients as *weights* (see Section 3.3) to determine the relatively more important CRA questions. In order to improve the accuracy of our model, we need to improve the weights that ensure only the relevant CRA questions are included in contract similarity calculations.

For variable weights, we use the KS statistics calculated for selecting the optimal window. Since the KS statistic is a measure of informativeness to predict the directionality, and it is automatically normalized within the range of [0,1] by definition, we simply use it as the variable weight. If the weight is 1 for a CRA question, the question is viewed as decisive when indicating directionality. If it is 0, there is no difference in the distributions between positive and negative Gross Profit Variance groups.

5. Enhanced Predictive Model Results based on Optimized Data Set Selection

Once the optimal data set and the new CRA question weights are determined, the next step is to test them with our KNN model. Thanks to the fully parameterized algorithm, it is possible to test over 900 threshold configurations and determine the most optimal values. With this enhanced version of the model, LLO-CV testing reveals significant improvement in all three metrics, as shown in Table 3. Our most important metric, NPCP, shows great improvement from 71% to 86%, while Directional accuracy and PCP also improve significantly from 59% to 76% and 52% to 68% respectively.

The improvement gained is due to two changes:

i) Data selection has been optimized. While the original model used the last CRA (E_1) and end of T&T KPI(ΔGP) (D_1) data for training, the new model uses the optimal windows E_2 and D_3 respectively.

ii) The CRA question importance weights have been improved as discussed in Section 4.3.1.

Table 3. Comparing initial and enhanced model accuracies

	Directional	NPCP	PCP	Engagmnt Training Data	Delivery Training Data	Run-time Window
Initial Model	59%	71%	52%	E_1	D_1	E_1
Enhanced Model	76%	86%	68%	E_2	D_3	E_2

An important consideration with this result is whether the model accuracy generalizes to other *run-time windows* (the E_*N* timeframe in which the prediction will be performed during Engagement) given that it is trained using only the optimal data set {E_2 X D_3}. We test this by running our optimally trained model on the test data obtained from all three run-time windows (E_1, E_2, E_3) in Engagement. The results, shown in Table 4, indicate that the accuracies obtained for optimal time run-time window (E_2) do not necessarily generalize to all run-time windows. While E_3 accuracies are very similar to optimal E_2, E_1 NPCP accuracy falls to 72%, well below the optimal 86% of E_2.

Table 4. Different run-time scenarios tested with optimally trained (E_2 and D_3) model

Run-time Window	Directional	NPCP	PCP	Engagement Training Data	Delivery Training Data
E_1	71%	72%	70%	E_2	D_3
E_2	76%	86%	68%	E_2	D_3
E_3	74%	81%	68%	E_2	D_3

We addressed this issue by splitting the model configuration into two settings: we use the optimal configuration (training data and thresholds) to train the model to be used in E_2 and E_3 run-times, and determine a new set of training data and thresholds that are optimal for E_1 run-time. Thanks to the automated training capability of our KNN based model, selecting and applying multiple configurations in real-time is trivial.

Table 5. Final model accuracies and associated optimal windows for training data

Run-time Window	Directional	NPCP	PCP	Engagement Training Data	Delivery Training Data
E_1	74%	75%	73%	E_1	E_3
E_2	76%	86%	68%	E_2	E_3
E_3	74%	81%	68%	E_2	E_3

To train the model that will run during E_1 run-time, we use a sub-set of the full data set that represented the E_1 timeframe (0-1 month before contract signing). Through the fully parameterized model, we also test the E_1 model with over 900 threshold configurations to determine the optimal thresholds. The final accuracies for E_1 run-time, shown in Table 5, are only slightly better than using the optimally trained model, with NPCP going from 72% to 75% accuracy. However, depending on the absolute number of contracts on which the model will be used, 3% improvement could be significant.

We should also note that, during the optimal threshold testing, the model yields a wide range of accuracies each of which is associated with the set of threshold values tested. The results shown are our selection of accuracies from this extensive set of results. Depending on the business goals, one could select a different result (and thus a different set of threshold values) to maximize NPCP, PCP or Directional metrics as needed. For example, if the goal is to maximize NPCP for E_1 run-time window, we could have chosen a threshold configuration with the 86% NPCP accuracy at the expense of 58% PCP accuracy. However, we selected the result reported in the top row of Table 5 (with NPCP at 75%), as we are looking for a more balanced model across all three metrics.

In summary, our final optimal model consists of two different parts, each of which is trained with its respective data set and optimal thresholds. In practice, when the risk managers or the QA experts perform a CRA, and want to use it to predict a contract's financial performance, the model trains itself automatically in real-time using the optimal data set and the optimal parameters of its run-time window. Such flexibility allows us to maintain optimal accuracy for our model as the training data set gets updated with new historical contracts over time.

6. Conclusions and Future Work

In this paper, we have described a methodology in building a financial performance prediction model with enhanced accuracy using ordinal risk assessment (survey score) data as model input. As a key part of our methodology, we have investigated how the time delay between contract risk assessments and contract performance measures within the IT service delivery lifecycle affects the accuracy of contract risk models in the IT outsourcing domain. We find that variations of the median time delay between contract risk assessments and the contract performance measures accounts for prediction accuracy variations as large as 25%. Moreover, we observe that statistical modeling strategies such as linear regression fall short when it comes to handling sequential and ordinal data sets, which are characteristics of the IT outsourcing domain. We show that, by using data mining and machine learning approaches, we can ensure selection of the optimal model parameters, thereby maximizing the accuracy of risk prediction models.

We conclude that the identification of relevant data selection criteria, such as the time delay between risk assessment and performance measurement, is key for optimizing prediction accuracy in data-driven, predictive risk modeling. Such optimized predictive models help enable proactive risk management and lead to cost reduction and improved quality in IT service delivery.

7. References

[1] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A. H.: Big data − The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute Report (2011).

[2] Gartner Business Process Mgmt Summit. London, U.K (2013).

[3] Taylor, J.: Predictive Analytics − Making Little Decisions with Big Data, In Information Management (2012).

[4] Güven, S., Tao, S. and Goma, S.: Financial Risk Analytics. In IFIP/IEEE International Symposium on Integrated Network Management, Ghent, Belgium (2011).

[5] Jan, E., Ni, J., Ge, N., Ayachitula, N., Zhang, X.: A Statistical Machine Learning Approach for Ticket Mining in IT Service Delivery. In IFIP/IEEE International Symposium on Integrated Network Management, Ghent, Belgium (2011).

[6] Abbott, G., Anerousis, N., Gordon, F., Grussing, A., Makogon, S., Manore, P., Humphries, F., Sherry, J., Tao, S.: Risk Identification and Project Health Prediction in IT Service Management. In IEEE/IFIP NOMS, Osaka, Japan (2010).

[7] Duda R., Hart P., Stork D.: Pattern Classification. 2nd Ed, Wiley (2001)

[8] Jain, A., and Zongker, D.: Feature Selection−Evaluation, Application, and Small Sample Performance. In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 9(2), pp. 153-158 (2002).

[9] Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection. In Journal of Machine Learning Research, vol. 3, pp. 1157-1182 (2003).

[10] William, R. L.: Effective Sample Size. In Encyclopedia of Survey Research Methods. In Lavrakas, P. J. (ed.), SAGE Publications, Inc, Thousand Oaks, California, USA (2008).

[11] Jin, R., Chen, W., and Sudjianto, A.: On Sequential Sampling for Global Metamodeling in Engineering Design. In Design Engineering Technical Conferences & Computers and Information in Engineering Conference (2002).

[12] Howell, D. C.: The Treatment of Missing Data. In the SAGE Handbook of Social Science Methodology. In Outhwaite, W., Turner, S. P. (eds.), pp 208-223, Sage Publications (2007).

[13] Witten, I., Frank, E.: Data Mining − Practical Machine Learning Tools and Techniques, Elsevier Publishing (2005).

[14] Stuart A., Ord K., Arnold S.: Kendall's Advanced Theory of Statistics. Volume 2A, Classical Inference and the Linear Model, 6th Edition, Wiley (2010).