Abstract

AI FOR CORPORATE INTEGRITY: IDENTIFYING PARTNER CORRUPTION RISK WITH DATA SCIENCE

Hyojin Song, AI for Good, hyojins@microsoft.com
Darren Tanner, AI for Good, datanner@microsoft.com
Moran Elnekave, Finance BI, melnekav@microsoft.com
Ryan McDonald, Finance BI, rymcdo@microsoft.com
Nicole Yoon, Finance BI, niyoon@microsoft.com
Rohini Kumar Madinadi, Finance BI, v-rohmad@microsoft.com
Kasi Pusuluri, Finance BI, kapusulu@microsoft.com
Joel Kalonji, CELA Data Science and Analytics, jokalonji@microsoft.com
Kenneth Chen, Finance BI, v-zhech@microsoft.com
Manjushree H. L, Finance BI, v-mahl@microsoft.com
Nitish Kumar Rai, Finance BI, v-nitr@microsoft.com
Puneeth Venugopal, Finance BI, v-puven@microsoft.com
Rahul Dodhia, AI for Good, radodhia@microsoft.com
Jeannine D'Amico Lemker, CELA Office of Legal Compliance, jdamico@microsoft.com
Alan Gibson, CELA Office of Legal Compliance, alangi@microsoft.com

Keywords. Compliance, corruption, business risk, anomaly detection, data science.

Microsoft brings its products to users through many channels, including via over 240,000 reseller partners. Although engaging end-customers through a global partner network brings numerous benefits, it also exposes Microsoft to numerous risks, including the possibility of engaging with corrupt partner organizations. Exposure to corruption risk through partners can create billions of dollars in fines, government imposed monitoring, and significant reputation damage. The goal of the High-Risk Partners (HRP) program is to protect Microsoft by identifying risky partners requireing additional compliance oversight through an early warning and monitoring system. Microsoft's previous partner vetting process relied on expensive external reports, did not make use of internal data sources, occurred in a once-yearly cycle, and was largely manual. Our team therefore built the HRP platform to directly score the riskiness of over 240k channel partners using internal data and on an ongoing basis. The HRP model performs with 44% precision for identifying high-risk partners, which contrasts with 12% precision for the previous manual approach. The HRP platform launched into production in January 2019 and is an essential compliance tool and key control for Microsoft. The platform provides human reviewers with actionable insights about partner corruption risk and creates a critical business impact when making partner vetting decisions.

1. Introduction

Microsoft's sales of software licenses and online services through its volume licensing programs generated over $71B in revenue in FY2019. Approximately 55% of volume licensing sales are brokered through third party channel partners, which include resellers, distributors, suppliers, and subcontractors. Although engaging end customers via channel partners brings numerous benefits, it also exposes Microsoft to risk, including potentially engaging with corrupt partner organizations. For example, Figure 1 depicts two hypothetical channel partner sales scenarios. The path on the right shows a transaction involving a single partner (e.g., reseller), who provides the product to the end customer. The path on the left involves multiple partners (e.g., reseller and distributor), as well as monetary concessions (e.g., discounts, end customer investment funds (ECIF)). Multi-layered transactions involve money changing hands more times with more intermediaries, thus obscuring the presence, timing, or locus of corrupt behavior in the process of getting Microsoft products to the end customer. Adding concessions increases opportunies for risk, where partners may not always pass on concessions such as discounts to the end customer, or where incentive funds can be used to bribe or provide kickbacks to public officials. In this context we can define partner corruption risk as encompassing bribery and incentive and concession abuse.

Indeed, Microsoft has faced investigations regarding scenarios just like this **(e.g., Fiscutean, 2014; Hinshaw & Greene, 2018; Matthews & Ovide, 2013**). Even with no wrongdoing on the part of Microsoft, exposure to risks like this can bring at minimum significant reputational damage, and conceivably lead to government-imposed monitoring, potential fines of tens (or hundreds) of millions of dollars, or even jail time for executives. This risk was recently highlighted by a settlement reached between Microsoft, the US Department of Justice, and the US Securities and Exchange commision worth $26M regarding just such a bribery scheme in the Hungarian subsidiary (Greene, 2019). Identifying problematic behavior and potentially corrupt partners early and systematically is obviously of paramount importance to the company both legally and financially.

Here we describe Microsoft's High Risk Partner (HRP) program, which is the first initiative to use data science to systematically identify potential corruption risk within Mirosoft's channel partner ecosystem. The HRP program provides a model-driven platform for human reviewers to identify and better understand corruption risk in partner organizations. It launched into production in January 2019 and is now an integral part of the company's partner vetting process when entering into and renewing contractual relationships with reseller partners. It augments the existing human-based vetting process managed by the OneVet organization, within Commercial Operations.

The HRP data model takes a hybrid rule-based analytical and data science-based anomaly detection approach to deliver a risk score between 0 and 100 for each partner. This hybrid approach is the result of collaborations between data scientists, engineers, and compliance experts; it was designed to provide more intuitive, actionable information than standard machine learning-only models can generally provide. This hybrid, interpretable structure has resulted in crucial buy-in from the organizations accountable for onboarding and offboarding partners from our channel ecosystem, which has in-turn led to broad adoption of the platform, with a resulting substantial business impact. The current scope of the HRP platform in production focuses on delivering insights for Microsoft's reseller partner network in the company's volume licensing space. However, the scoring logic is general and is being rolled out to other partner types (e.g., software advisors, cloud solution providers, etc.), as well as during other points during the partner lifecycle beyond vetting (e.g., when the company is approving discounts, investment funds, and credit term extensions).

Before describing and evaluating the HRP model, we first describe Microsoft's previous partner vetting approach, other machine learning work on identifying corruption risk,

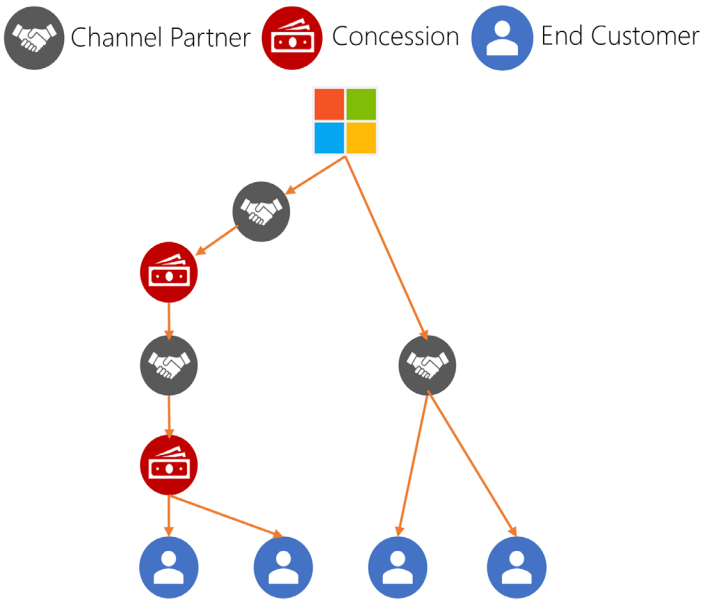

Figure 1: Schematic of channel partner sales model.

and finally traits of the HRP problem space and dataset that make our modeling target very different from prior work on corruption.

1.2. Previous Partner Vetting Approach

Partner vetting is carried out primarily by Microsoft's internal vetting organization, OneVet. OneVet performs numerous types of vetting for organizations across the company, including annual and, when necessary, on-demand partner corruption risk evaluations. However, most relevant to the present work is OneVet's annual enhanced anti-corruption review (EACR) cycle. This is the most in-depth type of vetting that occurs at the partner-level, and the outcomes of EACR serve as the only gold standard labels for partner corruption risk in Microsoft's network.

Each year, EACR is carried out for approximately 2000 partners. The EACR process can last weeks to several months for a given partner based on the presence and severity of corruption red flags found. The cost of the vetting process also varies by the level of depth needed, but the average monetary cost is approximately $1500/partner. Prior to the launch of the HRP platform, selection for EACR was carried out manually by Microsoft compliance experts based primarily on consideration of three factors: 1) the geographical area in which the partner is operating, which is mapped to area-specific corruption risk determinations provided by CELA in the form of the CELA risk tier; 2) the type of partner; and 3) the partner's revenue stream with Microsoft. Geographical area and revenue have traditionally been the largest factors driving selection. Because of these considerations, most reviews have historically focused on partners with large revenue streams operating in areas where corruption risk is known to be high.

EACR is a multi-tiered process, the exact details of which can vary based on the needs of a particular partner's case, but all reviews share some core elements. First, anti-corruption reports and scores are ordered from external providers for all partners selected for EACR. These reports detail adverse media and negative news, presence on specific watchlists, sanctioned parties, and politically exposed persons. Second, partners complete extensive questionnaires, with questions supplied by Microsoft's Office of Legal Compliance. This information, as well as information about prior review outcomes from previous EACR or OneVet reviews is taken to make an initial risk determination. For partners with no red flags for corruption risk at this review stage, approval takes days to weeks. Red flags in this case could refer to adverse media reports indicative of corruption found in the externally-sourced reports, or previous findings of corruption risk (see below), which may require follow-up diligence and risk mitigation (i.e., enhanced controls).

For partners with corruption red flags, more thorough vetting is carried out. The details of this vetting vary based on the exact context and type of red flag, but may include further research and diligence into partner business practices or validation of partner compliance with previously-implemented corruption controls. At this stage, partners with actionable corruption risks that can be mitigated by enhanced controls can be approved. In higher risk cases, a determination of "do not approve" is given and processes are started to restrict Microsoft's engagement with the partner. These cases are escalated to other compliance organizations within Microsoft, such as the Office of Legal Compliance, who do further work to determine the scope of disengagement. In certain high-risk cases, this could include fully off-boarding the partner from Microsoft's channel ecosystem.

Although this process has been successful in discovering and mitigating Microsoft's exposure to corruption risk in many cases, it does suffer from some drawbacks. First, vetting was only carried out once yearly. Even in cases where corruption risks may have become apparent through media reports about a particular partner at some point during the year, unless a specific on-demand vetting were requested, the media coverage would not be taken into account until the next yearly cycle. This cycle timing is not frequent enough to capture ongoing changes to partner risk profiles. Second, risk determinations were made primarily based on purchased external reports, and did not consider what in some cases are long-standing partnerships with Microsoft. In these instances, analysis of the partner's 360-degree business relationship with Microsoft could provide important clues to identify potential corruption risk. Third, decisions about inclusion in the annual cycle were made primarily based on two attributes: partner geography and revenue. Therefore, some partners with very low corruption risk but perhaps high revenue streams were vetted each year (i.e., redundant vetting despite no new evidence of risk), while some partners with smaller revenue streams but other potential red flags went unreviewed. Failing to asses risk for the entire partner ecosystem could have the unfortunate effect of exposing Microsoft to corruption risk that could otherwise be successfully mitigated.

1.3. Prior Modeling Approaches to Risk and Corruption

Risk modeling has been an area of extensive work within the machine learning (ML) literature. For example,

bankruptcy, credit default risk, and fraud detection (e.g., credit card or other transaction fraud) have received a large degree of attention **(e.g., Adewumi & Akinyelu, 2017; Barboza et al., 2017; Fu et al., 2016; Galindo & Tamayo, 2000; Hua, et al., 2007; Olson et al., 2012; Zheng et al., 2018)**, and more recently 'RegTech' for regulatory compliance in the financial industry has been a growing area of focus **(Aziz & Dowling, 2019; Wall, 2018)**. Bad transactions or regulatory breaches can often be identified with a clear binary label in supervised modeling (e.g., default or no default; regulatory breach or no regulatory breach). Current state of the art models typically adopt a binary classification approach using K-nearest neighbors, SVM, boosted trees, deep neural networks, Bayesian networks, or some stacked ensemble of multiple of these approaches.

In comparison to fraud or default risk, much less published work has focused on using ML to identify corruption risk. Corruption is more abstract than fraud or default making it harder to model, and acquiring labels requires lengthy (months- or years-long) investigations. As such, datasets are much smaller and labels are less reliable. Nonetheless, the few existing reports using ML to predict corruption have used traditional supervised approaches, with gradient boosted trees performing well in several applications **(Colonnelli et al., 2019; Gallego, et al., 2018; Grace et al., 2016)**.

One troubling issue in modeling corruption risk is that while outcomes of investigations may provide some clear positive (risky, corrupt) labels, negative labels are not always as clearly defined. Because of the cost and time overhead of launching corruption investigations, many datasets include large number of uninvestigated and therefore unlabeled cases. This leads to a non-trivial choice about what subset of cases to include in the ML model as negative labels. Including all cases, including uninvestigated and unlabeled cases, means potentially contaminating the 'good guy' pool in the model, since unknown bad actors may be among the unlabeled observations. Alternately, including only investigated cases in the model may drastically reduce size of the dataset. For example, when prototyping a model for predicting corruption risk in World Bank contracts, **Grace et al. (2016)** used only investigated cases, reducing their dataset size from ~200k total contracts to only 600. Such small datasets can lead to drastic overfitting, which is likely the case with Grace et al.'s model, and even make ML fully untenable.

1.4. Challenges in Modeling Partner Risk and Modeling Principles

Modeling corruption risk in the HRP space faces several challenges that make traditional supervised learning approaches like those describe above untenable. First, the models we use must be fully intelligible by human reviewers who are non-ML experts. Final risk determinations for partners are still made by OneVet reviewers. Goals for our models are to both indicate the riskiest partners for reviewers to focus their attention on, and to reveal to the reviewers those partners' riskiest attributes that require follow-up research and diligence, thus increasing efficiency and saving overhead costs. Because of this, no black box algorithms can be used. Even regression coefficients or outputs of tree interpreter algorithms (e.g., SHAP or LIME) can be difficult for non-experts to understand, especially when they involve marginal effects and interactions.

Second, only 2048 (<1%) of the 246k partners in our database have been reviewed; over 99% therefore have unknown risk profiles and no associated risk category label. Third, of the reviewed partners, only 87 (0.04% of the total dataset) received the highest risk assessment from OneVet (see Table 1 for risk category descriptions).

Category		Description	Risk Level	No. of Partners
1	Internal Issue Found: Escalated	Internally severe issues were detected by Audit or CELA OLC	Very High	87
2	OneVet Reviewed: Red Flag + AC Control	Vetted by OneVet; severe red flag found; anticorruption training was made as a control	High	165
3	OneVet Reviewed: Red Flag (No AC Control)	Vetted by OneVet; red flag found but not significant and thus no controls made	Med	95
4	OneVet Reviewed: No Red Flag	Vetted by OneVet; no red flags	Low/Med	1701
5	Not Reviewed	Not vetted	Unknown	>244K

Table 1: OneVet vetting outcomes with risk categories for partners in the dataset.

Taken together, this means that standard supervised learning methods that strictly classify a given partner as risky or not cannot be used. Restricting our dataset to include only investigated partners (i.e., partners with true labels) would make it too small for ML to yield reliable results. Additionally, the minority class (very high-risk partners) has so few instances that splitting the data into train/test strata and cross-validation folds would result in too few observations for any supervised learner to perform adequately, and there are too few instances for up-sampling methods such as SMOTE **(Chawla et al., 2002)** to perform well.

In some cases such as ours with incomplete labels, weakly- or semi-supervised learning approaches can be used to learn labels in addition to building a machine learning classifier **(see, e.g., Zhou, 2018)**. However, some of these approaches, such as active learning, assume a human 'oracle' who can adjudicate label predictions. In our case, such an oracle is unfeasible because of the cost (~$1.5k/partner) and duration of reviewing (months) by OneVet. Other approaches such as mixture modeling presume that data come from known probability distributions (e.g., Gaussian), which our data do not. Still other label-learning approaches assume that at least some labels are clearly defined as either positive or negative classes (e.g., high or low risk). However, as we describe in 3.1, below, we have no clear low-risk labels: the only labels we have are for partners who presented with at least some level of risk to human reviewers. Thus, even semi-supervised approaches cannot adequately meet the unique challenges that our data presents.

We therefore approached the present problem with four primary modeling principles in mind. First, the model cannot rely on labels for training, and as such, there is no objective cost function that can be minimized. Instead, we use a multiple anomaly detection ensemble approach. Second, because we do have reliable labels for a very small subset of our data (<1%) from prior human reviews, we can use this partial information about known bad actors to both determine optimal weights for ensembling individual risk models into an overall risk score and to assess model performance. Third, because final risk assessments are ultimately carried out by human compliance managers and because the aim of the program is to maximally facilitate these efforts, the model must maintain strict intelligibility. That is, no 'black box' algorithms can be used to derive risk profiles, and the model output must be fully transparent to and intelligible by human reviewers, who are subject matter experts in compliance, but non-ML-experts, via a dashboard tool. Fourth, we acknowledge that risk is not inherently categorical, but rather abstract and relative. Our models are therefore designed to produce relative risk scores for partners, with separate relative risk scores for specific risk areas.

2. Method

2.1. Modeling Approach

We approached these unique challenges by building a multiple anomaly detection ensemble model, which can intelligibly guide expert human reviewers for final risk determination. Our selected mechanism takes a multi-stage approach by breaking the overall model into a set of component models. First, we obtain risk scores for each attribute (e.g., revenue, discounting, ECIF, etc.). This score is in-turn made up of subscores that consider different aspects of the attribute, with scoring approaches appropriate for each respective attribute and its corresponding data type (see Section 2.2, below). Each risk attribute score is scaled between 0 and 100, representing the low and high ends of the risk spectrum, respectively. We then built a final risk score from the attribute scores using a weighted linear average. Weights were determined using the partial information about partner risk available from prior reviews (see Section 3.2 for more detail), combined with input from compliance experts and stakeholders from the compliance reviewer community. This multi-stage approach maintains interpretability for non-ML-expert human reviewers so that the largest risk attribute contributor(s) to a partner's final risk score can be communicated to the reviewer, who can then direct their review efforts to the areas that present the most pressing concerns for compliance risk.

In this way our multi-stage approach bears some resemblance to model stacking methods described in previous literature **(e.g., Breiman, 2004; Leblanc & Tibshirani, 1996; Sill et al., 1999; Smyth & Wolpert, 1999; Wolpert, 1992)**. However, in standard stacking approaches, model predictions are generated by pooling the predictions of several independently-trained learners of the same type (e.g., regression, neural networks), which are each trained for the same objective but with a different subset of data. Our approach bears more resemblance to feature-weighted linear stacking **(Sill et al., 1999)** than to pure stacking methods; however, even feature-weighted stacking methods generally combine not only multiple feature-weighted instances of a single learner type, but also pool multiple different types of learners, each of which is aimed at the same outcome (e.g., multiple weighted instances of both singular value decomposition and k-nearest neighbors for

recommender engines, with weights for each contributing model determined by the value of some feature). Our component models are each trained on separate datasets with separate modeling objectives, and then linearly combined into a single risk score assignment.

2.2. Risk Areas and Sub-Models

Major risk areas that are included in the HRP model risk score are described in the appendix, in Table A1, along with information about the respective risk source. Numerical contributions of these risk areas to the overall score are depicted in Figure 2. We first group risk attributes into three broad risk areas: entity trustworthiness, 360° business relationship, and business environment. Scores for entity trustworthiness are provided to us by other vetting organizations within Microsoft when available, with scores being determined by that organization based on its own research and due diligence. These scores are based on the external reports that were previously the primary determinants of risk for partner vetting. Scores in the other two major risk areas – 360° business relationships and business environment – are computed in our model from Microsoft-internal data. CELA risk tier information aside, these two areas and their component sub-models represent the major novel contribution of the HRP program to Microsoft's vetting process.

The 360° relationship score is comprised of risk scores derived from previously found issues, partner revenue data, and data on payments made to partners. Each of these is further built from a set of risk attribute scores (terminal nodes in Figure 2). The score for each risk attribute is further derived from one or more models. Where there is more than one model, the various models consider different aspects of the risk attribute and score in slightly different ways (see Tables A2-4 for details). For most models, scores for a given

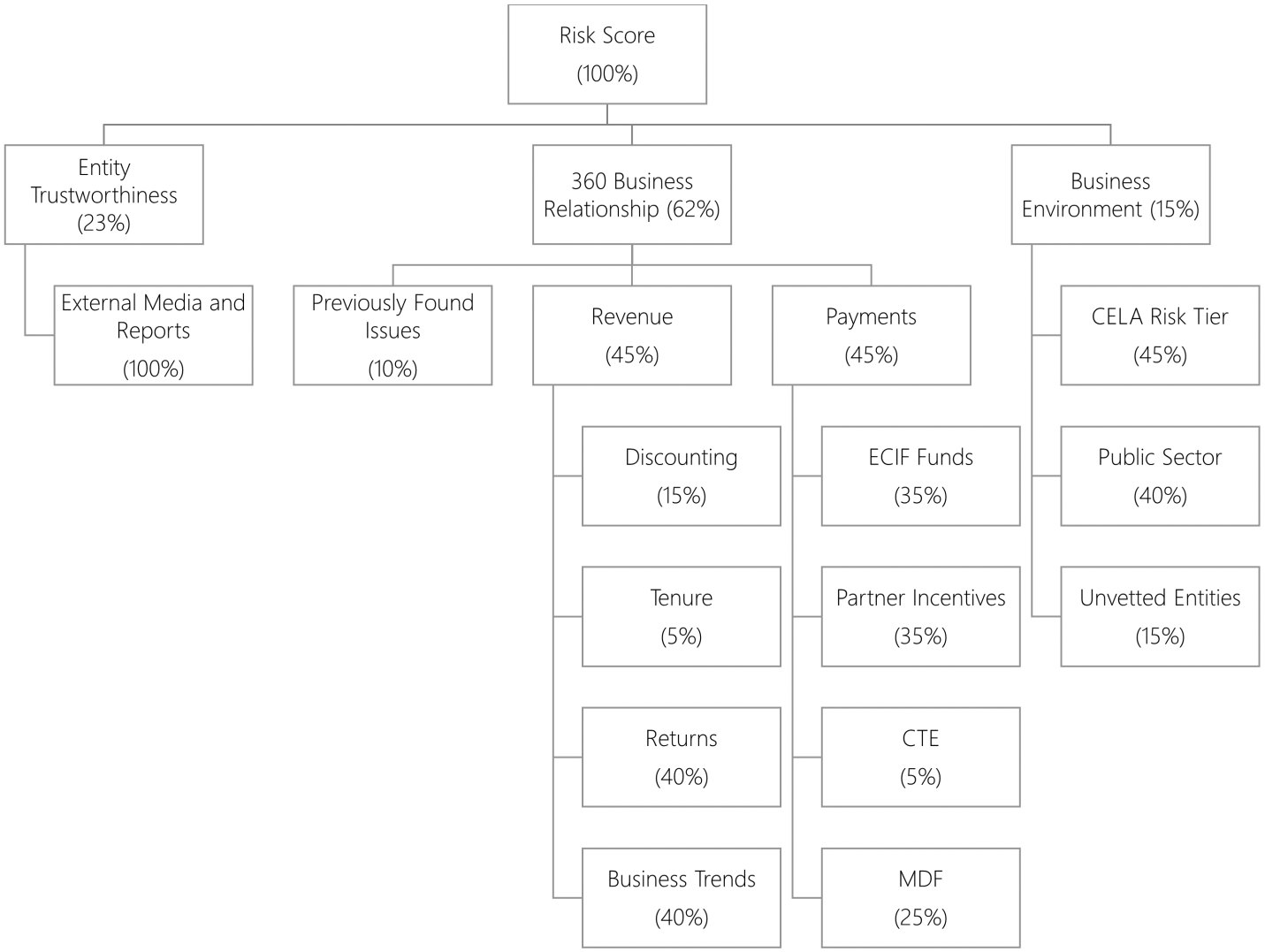

Figure 2: Major risk areas and contributions to final risk score. Score for a given node reflect that node's contribution to the parent's score.

partner are assigned based on the partner's location in peer comparison distributions, where peer groups are defined based on partner area, region, sub-region, subsidiary, and sector. The targets of each risk attribute and sub-model, as well as the scoring logic/method for each sub-model, were determined based on consultation with domain experts in risk and compliance.

Most models are built around a logic of ranking partners relative to peers, taking into consideration different time frames. Some scores focus on the most recent financial quarter; some consider the past eight quarters in aggregate; some consider each of the last eight quarters individually, and then assign an overall score by taking the maximum risk score across quarters and peer groups (when a given partner can be assigned to more than one peer group). More details about each of the component risk scores can be seen in the appendix, in Tables A2, A3, and A4.

In these tables, 'hockey stick' scoring functions refer to a scoring logic where partners in the lower tier of percentiles of risk in their peer comparison receive a score of 0, but above an inflection point (e.g., above the 85th percentile), partners receive increasing scores, which are linearly scaled between 0 and 100. The 'stretched Z' function describes a scoring logic where partners whose percentile rank falls below an inflection point (e.g., <70th percentile) all receive scores of 0, partners within a specified percentile interval (e.g., 70th >= partner percentile < 85th) receive scores linearly scaled between 0 and 100, and partners above the interval all receive scores of 100. Other models assign specific scores to bins of partners based on a scoring logic determined in consultation with domain experts.

One particular type of ML anomaly detection model used as part of the risk attribute score for four of the payments attributes is the Isolation Forest. Isolation Forests **(Liu, Ting, & Zhou, 2008, 2012)** are an unsupervised, nonparametric anomaly detection algorithm suitable for large datasets, where multiple feature dimensions need to be considered simultaneously in order to identify multivariate outliers. A given isolation tree is created by recursively partitioning feature space; forests are created by ensembling multiple trees in an analogous way to the better-known random forest algorithm. However, unlike random forests, the feature chosen to split over and the split point are randomly selected. That is, no objective cost function is being minimized, and the model does not 'learn' in the traditional sense. Isolation Forests are a fully unsupervised partitioning algorithm that can be used to generate anomaly scores.

3. Results

3.1. Component Score Evaluation

As mentioned above, risk category labels are available for <1% of partners. Although this is too few to use for traditional supervised machine learning, we can use these known risk labels for this subset to assess the performance of individual component (risk attribute-level) models. Risk category labels were derived from OneVet review outcomes and are described in Table 1; for reviewed partners, these are true gold standard labels. Five categories are delineated. Over 99% of partners fall into Category 5 (Not Reviewed) and therefore have unknown risk. Presumably, the vast majority of these partners pose no corruption risk; however, the lack of any vetting information on them does not allow us to clearly categorize them as such. Partners in Categories 3 and 4 were manually selected for review based on their partner profile, which indicates some degree of perceived risk in their portfolio (primarily large revenue streams and/or business in a geographical area with known high risk). However, further investigation by OneVet raised no concerns (Cat. 4) or previous red flags, which have been successfully mitigated (Cat. 3). We thus classify these partners as low/medium and medium risk, respectively. OneVet reviews for partners in Categories 1 and 2 did reveal significant compliance issues, which led to either anti-corruption mitigation controls (Cat. 2) or further escalation within Microsoft's compliance ecosystem, such as to the Office of Legal Compliance, One Commercial Partner (OCP) organization, or Partner Audit program in Internal Audit (Cat. 1). In the case of Category 1 partners, escalation outcomes could potentially lead to partner off-boarding. We consider these to be high- and very high-risk partners, respectively.

If our individual anomaly detection models are sensitive to partner-level risk, we should observe a correlation between risk category and anomaly score (i.e., higher anomaly scores for Cats. 4 and 5, than for 1 and 2). Importantly however, we do not expect any one scoring model to uniquely or exhaustively discriminate between lower and higher risk partners. Rather, the trend across partners should show higher score assignment to higher risk partners, and this should generally hold across risk attributes.

Here we show evaluations of four separate risk attributes: ECIF, discounting, partner incentives, and public sector business. Where there is more than one model constituting a risk attribute in Tables A2, A3, and A4, the plots show the weighted average of those models, combined based on

the weightings in the corresponding table (the procedure for determining these weights is described in 3.2).

Depictions of mean risk scores and 95% confidence intervals (CIs) for these risk attributes as a function of risk category are shown in Figure 3. We see that average partner risk scores increase with risk category for all four of the risk attributes, though the overall value of the mean risk scores differs across attributes. Average values are higher for discounting and public sector business than for ECIF and partner incentives. This is because very few partners actually receive ECIF or incentive payments. Because risk is only found when these payments are distributed to partners, partners not receiving these funds are given scores of 0 for the corresponding risk attribute. That is, if Microsoft does not provide the partner with ECIF or incentives, the partner cannot abuse these payments, so no risk is present. Thus, the lower mean scores reflect a large number of zero values contributing to the bins' averages. Note also that the 95% CIs of the mean show little overlap across categories. This is the case in particular for the highest risk categories (Cats. 1 and 2), as well as for the unvetted and lowest vetted risk categories (Cats. 5 and 4, respectively).

These results show that, on average, our anomaly-based risk scores are useful for identifying risky partners in general. Moreover, by computing separate risk scores for separate sources of risk (i.e., first scoring within distinct risk attributes), we can provide individual risk attribute scores to OneVet reviewers. This then allows the reviewers to focus their research and vetting efforts on the riskiest and potentially most problematic areas of a partner's business milieu and Microsoft's business relationship with the partner.

3.2. Final Risk Model Composition and Results

After identifying risk attributes and building attribute-level scores, weights for the final ensemble risk score were determined. Our process for weight optimization departed from traditional ML practices due to a series of practical constraints. First, because of the modeling challenges we faced (Section 1.4) traditional performance measures to maximize via cross-validation such as precision, recall, and F-beta scores could not be used, as the delineation between positive (risky) and negative (safe) cases in our data is not precise. That is, only a small number of bad actors who were offboarded or escalated via CELA OLC, OCP, and Audit are known, and this list is incomplete, since the vast majority

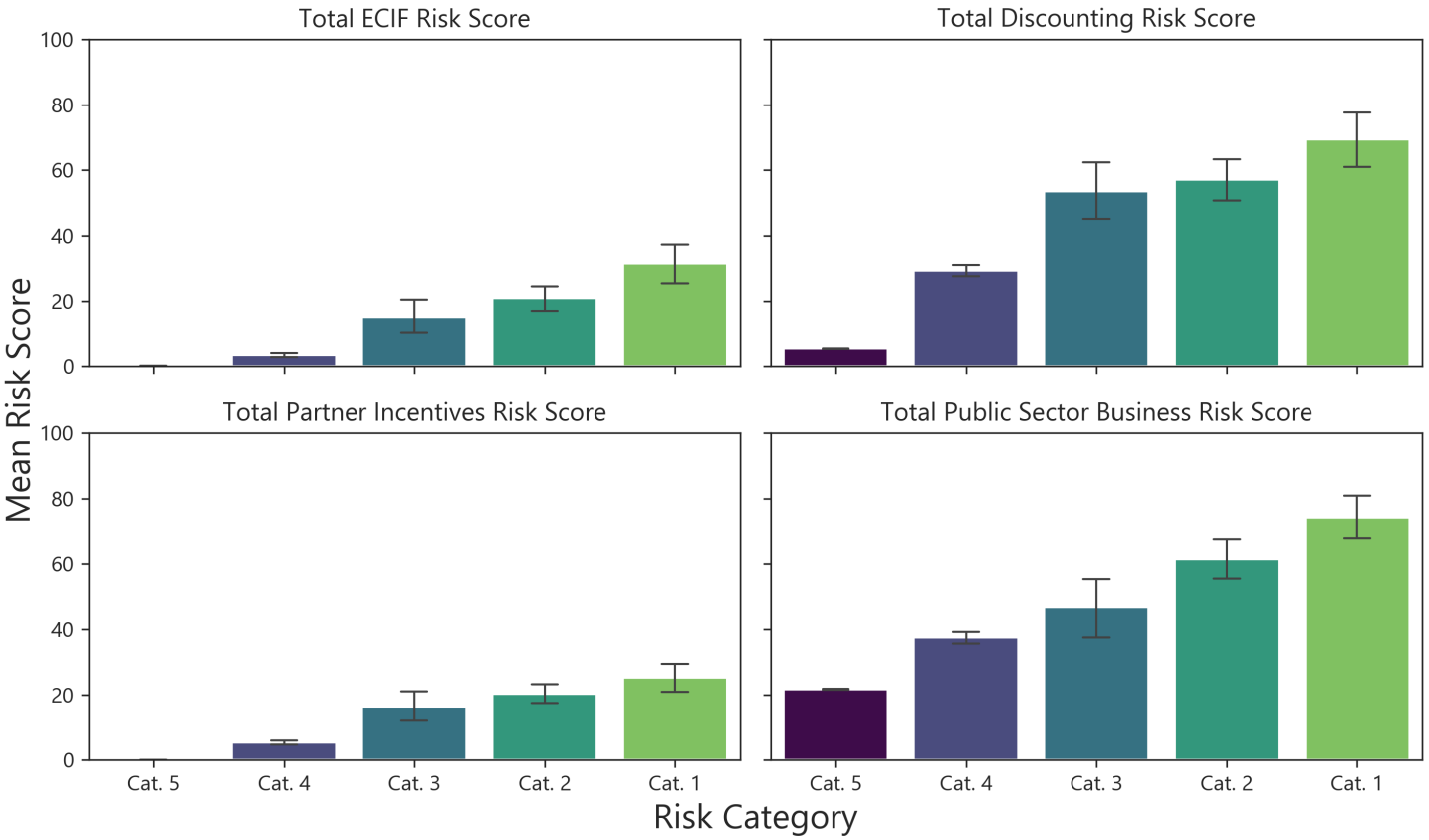

Figure 3: Mean risk scores for ECIF, discounting, partner incentives, and public sector business as a function of risk category. Risk categories are those from Table 1. Error bars show 95% CIs of the mean.

of partners did not receive any form of OneVet review. Moreover, the list of clear good actors among un-reviewed partners is not known. Thus, we cannot compare our predicted outcome or risk score with any true state for the vast majority of partners. Furthermore, because the number of known Category 1 (highest risk) partners is exceedingly small, performing train/test splits and cross-validation to optimize feature weights for the level-2 ensemble model was untenable. Finally, a key component necessary for the success of the HRP platform is buy-in from the reviewer community; without trust and endorsement from the reviewers who are the intended users of the product, some of whom are skeptical of 'black box' machine learning scores, insights our platform provides might go unactioned. Input from the reviewer community on how scores should be combined was therefore taken into consideration.

With these constraints in mind, we used a two-stage optimization process. The first stage involved quantitative optimization; the second stage involved human adjustment to the weights determined in the first stage. In the first stage, risk scores were computed for each partner over a series of weight combinations using a grid search approach. We then identified the proportion of Category 1 (very high-risk) partners captured in the top 1%, 5%, and 10% of risk scores for each weight combination. Identifying the weight combinations that maximized this proportion, which we term 'pseudo-recall,' was the primary objective in this stage, and provided a set of candidate weights for the second stage.

We chose the name 'pseudo-recall' for the optimization objective because our metric is related to recall, but true positive and false negative rates cannot be calculated exactly. Unlike binary classification models, our model does not return the probability of a partner belonging to a given class, but instead gives an overall risk score; the objective of the risk score is not discrete categorization based on class membership probabilities. Additionally, the number of false negatives cannot be calculated, since the majority of partners remain unvetted. Instead, pseudo-recall computes the proportion of known very high-risk partners who fall above a given quantile threshold out of the population of known very high-risk partners. We strove to maximize pseudo-recall because our primary objective is to identify all known risky partners, so as to minimize Microsoft's potential exposure to said risk.

After obtaining a set of weight combinations that maximized pseudo-recall at the three different percentile cut-off values, they were combined into a single weight combined into using human evaluation of their commonalities; these weights were then human-adjusted for the final model after considering a set of desiderata. These included maintaining high pseudo-recall at the 1%, 5%, and 10% levels, while also choosing a combination that would be transparent to, and interpretable and trusted by, the users of the HRP platform (OneVet reviewers). The first way this was achieved was by reviewing the set of highest-performing weights, and constraining risk attribute-level weights for the *a priori* chosen risk attributes (described above) to be non-zero. Next, component score weights were rounded to the nearest 5% for individual risk attributes, based on how they were to be reported in the final reviewer dashboard. This constraint was added to aid interpretability to reviewers. Where weights in a given risk attribute in Tables A1, A2, and A3 are not round, those individual scores are displayed as a single summed score in the reviewer dashboard (e.g., PI1, PI2, and PI3 in A3 are reported as a single score totaling 40% of the partner incentives attribute score in the HRP dashboard).

Pseudo-recall metrics for the final weights are given in Table 6, and mean risk scores by risk category with 95% CIs for the final model are depicted in Figure 4. When considering approximately the top 1% of partner risk scores, 81.4% of known risky partners are captured. When considering the top 5.7% of risk scores, pseudo-recall rises to approximately 92%. Figure 4 shows that mean scores increase monotonically across increasing risk categories, and 95% CIs for the lowest

Top quantile	No. Risk Cat. 1 Partners Flagged	Risk Cat. 1 Pseudo-recall
0.15%	64	65.98%
1.12%	79	81.44%
5.72%	89	91.75%
9.75%	96	98.97%
16.24%	97	100%

Table 6: Final risk model pseudo-recall scores.

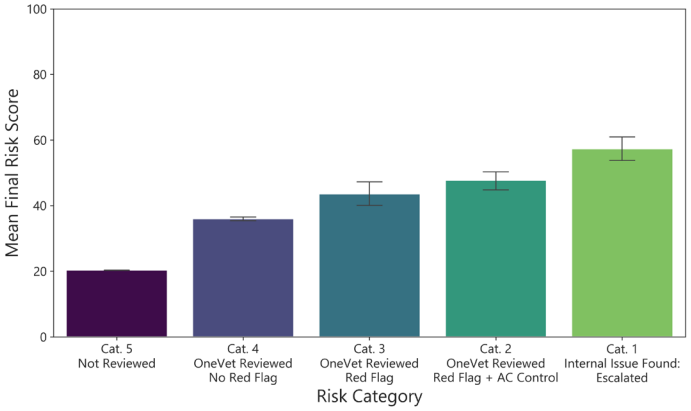

Figure 4: Mean final HRP risk score by risk category. Error bars show 95% CIs of the mean.

risk categories (4 and 5) show no overlap with the highest risk categories (1 and 2).

The final model therefore shows good separation in scores between the lowest and highest risk partners, among those partners in our database during model construction. However, unlike traditional ML approaches where the final model is validated on a hold-out set of the data, this evaluation was carried out on the dataset from which the model was built. In principle this can lead to overfitting of the development data, and poor generalization to new data. However, as discussed above, because of the extreme imbalance in our data and the paucity of known high-risk partners in our existing partner ecosystem, using a traditional hold-out dataset was untenable for the present problem. This initial evaluation should therefore be considered exploratory. A true test of the efficacy of our model requires evaluation on a novel dataset over which weights were not optimized. Such a dataset first became available after our model was launched in production, and it will be discussed below.

3.3. Production and Ecological Validation

The High Risk Partners Platform was launched in January 2019, and it is in use by OneVet for partner review in the volume licensing space. In the OneVet platform, the risk score for a given partner is presented for reviewers in a dashboard along with the top-5 risk drivers for a given partner. On a partner's landing page, reviewers see the final risk score given by the model, along with information that led to that specific risk score determination. The dashboard provides a high-level overview of a partner's risk score components (Figure 5), with qualitative relative risk information by risk attribute (high/medium/low risk flags based on final risk score bins). Reviewers can then click into any risk attribute for a deep dive into the data and business drivers behind any risk attribute score with embedded PowerBI (Figure 6, next page). This dashboard thus provides reviewers not only with the final score for a partner, but also information on how that score was derived. This includes both information on which risk attributes are driving the score, as well as the ability to dive into the raw data supporting the score.

Since the platform was launched, we have acquired a large set of new labels obtained on an out of sample dataset in the form of Cloud Solution Provider (CSP) vetting outcomes. This partner type is not normally included in the annual EACR cycle. Because CSP partners are relatively new in Microsoft's channel partner ecosystem, and because the CSP space represents one of the company's top areas for growth, OneVet is currently engaged in a special reviewing cycle to rapidly vet and assess risk among CSP partners, with the goal of vetting all partners new to the CSP ecosystem. The present dataset reflects the partner vetting population that was completed during the second half of FY2019. These partners were not included in the dataset used to optimize model weights, and therefore represent an ideal out of sample dataset. Importantly, this dataset is not a traditional hold-out dataset, which may not necessarily reflect a good proxy for the types of data that the model will be used for once

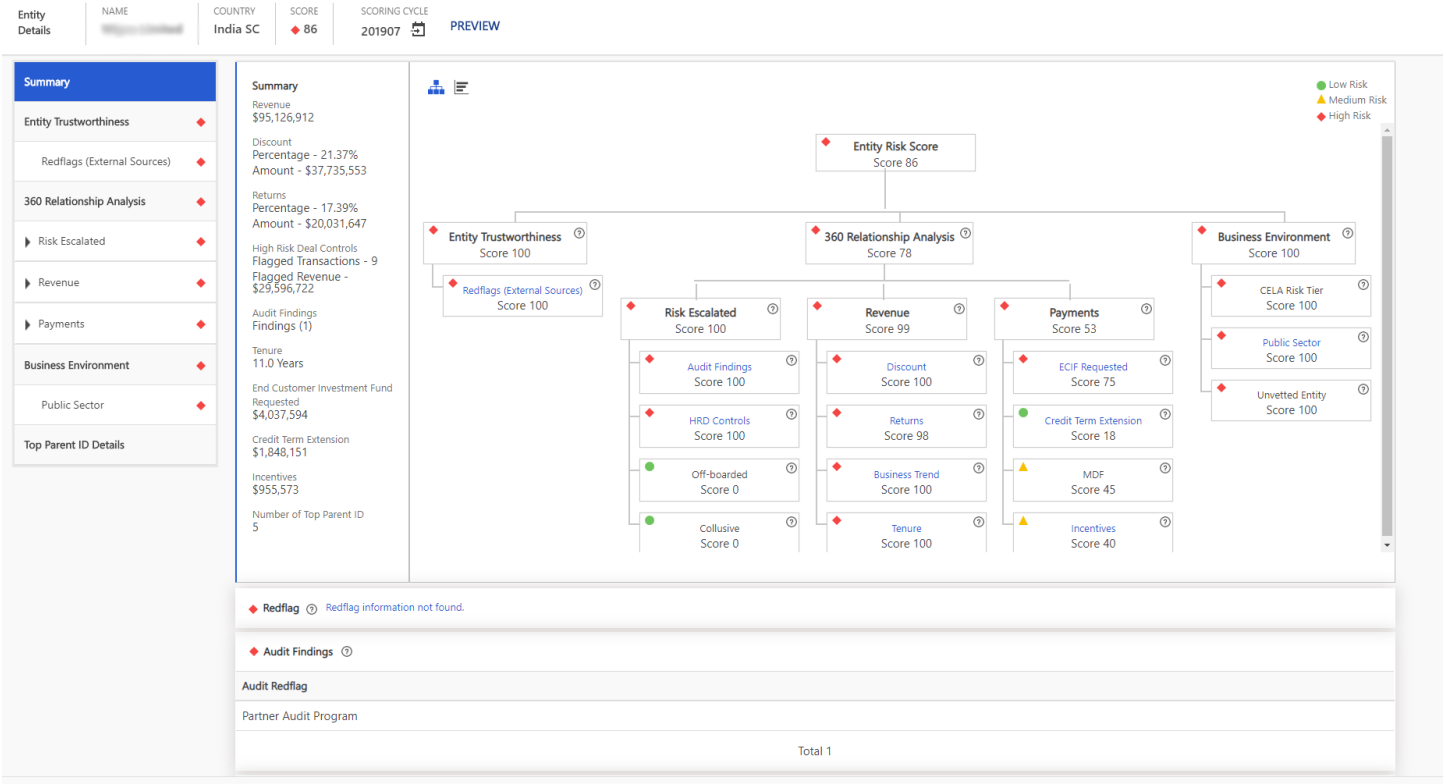

Figure 5: High level risk score overview from HRP dashboard for hypothetical partner.

deployed. Rather, this is a true ecological validation to assess the efficacy of HRP scores in predicting risk in real-life scenarios.

During the second half of FY2019, vetting outcomes were received for 8542 CSP partners. From the OneVet determinations, we categorized partners as either high-risk or low-risk. Partners categorized as high-risk are those who were rejected (and escalated), approved with anti-corruption controls because of red flags, or assigned to high-risk tier determinations for further compliance monitoring; partners who were approved without any controls were categorized as low risk. 344 partners were determined to be high-risk, and 8198 were determined to be low risk.

Density-normalized HRP risk score distributions for the high and low risk partners are depicted in Figure 7. As can be seen, the high and low risk partners appear as separate distributions. That is, the plot shows that HRP scores provide a very good, though not perfect, predictor of partner corruption risk. A Welch's t-test for independent samples (assuming unequal variance) showed that the mean scores for the two partner types were reliably different ($t = 26.02$, $p < .00001$).

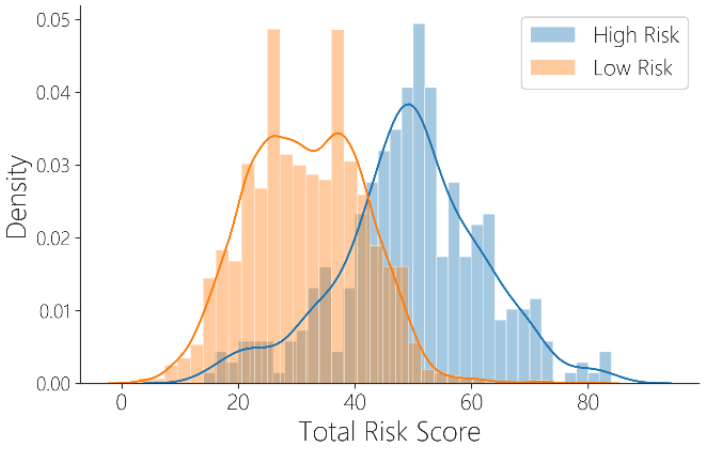

Figure 7: Density-normalized HRP risk score distributions for high and low risk CSP partners.

Currently in the HRP platform, partners who receive a total score of 50 or greater are flagged as 'high-risk partners' and recommended for vetting review. We refer to these partners as being "system flagged." To asess the HRP model's ability to disciminate high and low risk partners, we used system flagging as a binary categorization variable, and compared these flags against the high and low risk determinations by OneVet (that is, the ground truth determinations). A confusion matrix with raw and column-normalized values is presented in Table 7. Using system flagging with a threshold score of 50 leads to a recall of 51% and precision of 44%.

		System flagged?			
		Raw		Column-normalized	
		No	Yes	No	Yes
Vetting Outcome	High Risk	169	175	0.021	0.439
	Low Risk	7974	224	0.980	0.561

Table 7: Confusion matrix for CSP vetting validation outcomes.

Program	FY18-Q1	FY18-Q2	FY18-Q3	FY18-Q4	FY19-Q1	FY19-Q2	FY19-Q3	FY19-Q4	Total
EA	18.89%	16.03%	23.99%	17.98%	15.83%	14.69%	22.27%	45.36%	21.02%
Open	0%	0%	0%	0%	0%	0%	0%	0%	0%
Other	0%	0%	0%	0%	0%	0%	0%	0%	0%
Select	0.07%	0.21%	0.21%	0.3%	0%	0.07%	0%	0%	0.18%

Sector	FY18-Q1	FY18-Q2	FY18-Q3	FY18-Q4	FY19-Q1	FY19-Q2	FY19-Q3	FY19-Q4	Total
Commercial	17.58%	15.17%	23.92%	17.07%	16.73%	14.9%	22.31%	100%	21.02%
Public Sector	25.57%	0%	5.26%	23.76%	0%	0%	0%	0%	19.8%

Figure 6: Discounting risk attribute score deep dive for hypothetical partner in HRP dashboard.

To compare this outcome with the prior OneVet selection process, if we take the ~2k partners selected for vetting as a "predicted positive," and binarize the partner risk categories delineated in Table 1 into high and low risk (where Cats. 1 and 2 are considered high-risk and Cats. 3 and 4 are considered low risk), we can compute a precision of 12%. True recall rates for the prior OneVet selection process cannot be computed due to the issues discussed in Section 3.2. This precision rate means that only 12% of partners selected for review using OneVet's prior selection strategy actually showed high-risk attributes needing mitigation or escalation. This contrasts with a 44% hit rate for high-risk partners among those who would be selected for review based on our system flagging threshold from the HRP score.

4. Conclusions and Future Directions

The High Risk Partners platform is currently one of the most impactful compliance tools in Microsoft. It is the product of collaborations involving numerous stakeholders across disparate parts of the company, and now provides important, actionable insights to OneVet reviewers who make final risk determinations during partner vetting. It provides an early monitoring and warning tool for Microsoft's compliance and corruption monitoring organizations, by giving humans both quantitative and qualitative insights in an interpretable and user-friendly manner, while still maintaining scoring integrity and ability to identify partners in need of further vetting scrutiny.

The current product is in use by OneVet and is now a key driver for partner vetting selection in the volume licensing reseller space. The platform is currently being rolled out to include CSP resellers, and further extensions to include partner types, including software advisors, distributors, and others, are in progress. The solution we described here overcomes some of the key limitations of OneVet's previous partner selection and vetting system described in Section 1.2. Namely, partners are recommended for vetting based on anomalous behavior across a number of data streams available from Microsoft-internal data, and not simply on partner geography and revenue. Moreover, partner scoring now occurs quarterly, as new quarterly financial data are ingested, with a goal of moving toward continuous scoring. In this way, changes in partner details that can be indicative of risk can be taken into account as data becomes available, and not simply during the once-yearly standard EACR vetting cycle. The platform additionally scores all partners for risk simultaneously, and therefore provides a window into risk profiles for smaller partners and partners from non-risky geographical areas, who would have gone unscrutinized under the previous system.

Throughout the model building process we faced a number of data challenges and balanced a number of competing constraints. Most relevant to the data science and modeling problem were the lack of reliable labels for the vast majority of our dataset, as well as the severe imbalance among those partners who were labeled. This made traditional ML methods and solutions such as supervised learning, or validation methods and metrics untenable in the HRP program. Additionally, we weighed the concerns, opinions, and needs of the human reviewers who use our product heavily in both the model building process and design of the final dashboard tool, where the model outputs are reported and described to the reviewer community. Without trust and buy-in from reviewers, the production model would go unused in practice. The final model approach and architecture therefore reflect a balance of data-driven optimization constraints with human-centered design and reporting considerations.

With this initial model in place, the HRP program will be able to rapidly iterate, grow, and mature our data science approach in at least two major ways. First, the introduction of a model-based system to the reviewing community has the effect of familiarizing the reviewers with data science and machine learning concepts, and will build trust among reviewers in the accuracy and efficacy of data science-based models. Once a base level of trust is established, more sophisticated (and less transparent) ML models will be able to be introduced without sacrificing buy-in among the user base.

Second, instead of manually selecting partners for review out of an unsorted heap, our HRP scores will enable reviewers to focus their vetting efforts on only those partners who have profiles strongly indicative of corruption and compliance risk. This will speed the identification of risky partners, who may have been missed by the less structured previous selection approach. Metaphorically, instead of looking for risky needles in a haystack, reviewers will be looking for needles in a handful of hay. In turn, this will quicken the accumulation of high-risk partner labels. As these labels are added to our dataset, more accurate supervised learning approaches to corruption risk will become tenable.

5. Appendix

Risk Area	Risk Rationale
Entity trustworthiness	Scores from external vetting groups using media and other reports on partner organization trustworthiness indicating partner risk.
Previously found issues	Partners who have prior partner- or contract-level red flags from vetting outcomes or who have previously been off-boarded are risky.
Discounting	Partners receiving larger discounts relative to peers are risky.
Returns	Partners who return products at higher rates and for higher amounts relative to peers and relative to their revenue are risky.
Business trends	Partners with unusually large deals relative to peers are risky.
Business tenure	New partners with unknown track records are risky; extremely long-tenure partners present risk through high knowledge of concession system.
End Customer Investment Funds (ECIF)	ECIF is payments for services delivered to end customers in support of Microsoft products, to drive deployment or migration of products, or to provide customer support. Large payments relative to peers are risky.
Credit Term Extensions (CTE)	Partners requesting large or long extensions to credit terms relative to peers are risky.
Marketing Development Funds (MDF)	MDF is payments made to partners for marketing purposes. Large payments relative to peers are risky.
Partner incentives	Disproportionately large incentive payments, or an unusual number of payments, made to a partner relative to revenue/peers are risky.
CELA Risk Tier	Risk tier of geographic location of partner assigned by Microsoft's Corporate, External and Legal Affairs group. Partners doing business in areas with higher known rates of corruption are riskier.
Public sector business	Partners doing business in areas with a high proportion of business with public sector business are riskier.
Unvetted entities	Partners in geographic subsidiaries where few partners have been vetted are potentially riskier.

Table A1: HRP risk attribute definitions and risk rationales.

Risk Attribute	Model	Weight (per risk attribute)	Scoring Overview
Discounting	DISC1	40.0%	Compute z-scores for average discount over last 8 quarters within peer groups; score using stretched Z function.
	DISC2	60.0%	Angle for ratio of discount amount (as percentage of total sales) to total sales using arctangent; angles scaled between 0 and 100.
Tenure	TEN	100.0%	Compute percentile ranks for partner tenure within peer groups. Short tenures receive decreasing scores; intermediate tenures score 0; long tenures receiving increasing scores.
Returns	RET1	5.6%	Compute percentiles for partner returns as percentage of transactions within peer groups over last 8 quarters; score with hockey stick function.
	RET2	22.4%	Compute percentiles for partner returns as percent of total partner revenue within peer groups over last 8 quarters; score with hockey stick function.
	RET3	2.4%	Compute percentiles for number of returns in the most recent fiscal quarter within peer groups; score with hockey stick function.
	RET4	9.6%	Compute percentiles for total monetary value of returns in last fiscal quarter; score with hockey stick function.
	RET5	60.0%	Angle for ratio of returns amount (as percentage of total sales) to total sales using arctangent; scale angles between 0 and 100.
Business Trends	BT1	40.0%	Compute percentiles for total contract value per partner within peer groups over last 8 quarters; score with hockey stick function.
	BT2	60.0%	Compute percentiles for total contract value per partner within peer groups for each quarter; score with hockey stick function; take max over previous 8 quarters.

Table A2: Scoring description for models contributing to Revenue score.

Risk Attribute	Model	Weight (per risk attribute)	Scoring Overview
ECIF	ECIF1	10.0%	Calculate ECIF amount as percent of total revenue over previous 8 quarter; penalize high percentages.
	ECIF2	10.0%	Calculate z-scores for ECIF amount per geographical area over previous 8 quarters; penalize z-scores above 1.
	ECIF3	10.0%	Calculate number of projects receiving ECIF funds per partner per geographical area per quarter; penalize high counts.
	ECIF4	10.0%	Calculate number of ECIF-receiving projects for each partner-customer dyad; penalize high counts.
	ECIF5	60.0%	Isolation Forest anomaly detection score.
Partner Incentives	PI1	13.3%	Calculate z-scores for partner incentives received per quarter. Penalize z-scores above 1.
	PI2	13.3%	Calculate percentage increase in partner incentives going into most recent quarter; penalize high percentages.
	PI3	13.3%	Calculate number of times partner received incentives in a given quarter over previous 8 quarters; penalize high counts.
	PI4	60.0%	Isolation Forest anomaly detection score.
Credit Term Extensions	CTE1	10.0%	Penalize long-duration CTEs.
	CTE2	10.0%	Penalize high number of CTE enrollments.
	CTE3	10.0%	Calculate z-scores for CTE amounts; penalize z-scores over 1.
	CTE4	10.0%	Penalize CTE amounts greater than geo area average.
	CTE5	60.0%	Isolation Forest anomaly detection score.
Marketing Development Funds	MDF1	20.0%	Calculate z-scores for MDF amounts per geo area; penalize z-scores above 1.
	MDF1	20.0%	Calculate z-scores for number of distinct MDF POs received in a given geo area; penalize z-scores above 1.
	MDF2	60.0%	Isolation Forest anomaly detection score.

Table A3: Scoring description for models contributing to Payments score.

References

Adewumi, A. O., & Akinyelu, A. A. (2017). A survey of machine-learning and nature-inspired based credit card fraud detection techniques. *International Journal of Systems Assurance Engineering and Management, 8(s2)*, 937–953. https://doi.org/10.1007/s13198-016-0551-y

Aziz, S., & Dowling, M. (2019). Machine learning and AI for risk management. In T. Lynn, J. G. Mooney, P. Rosati, & M. Cummins (Eds.), *Disrupting Finance: FinTech and Strategy in the 21st Century* (pp. 33–50). https://doi.org/10.1007/978-3-030-02330-0

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications,* 83, 405–417. https://doi.org/10.1016/j.eswa.2017.04.006

Breiman, L. (2004). Stacked regressions. *Machine Learning,* 24(1), 49–64. https://doi.org/10.1007/bf00117832

Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research,* 16, 321–357. https://doi.org/https://doi.org/10.1613/jair.953

Colonnelli, E., Gallego, J. A., & Prem, M. (2019). What predicts corruption? *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.3330651

Fiscutean, A. (2014, October 2). Romania opens corruption case into nine ministers over Microsoft licensing deal. *ZDNet. Retrieved* from https://www.zdnet.com/article/romania-opens-corruption-case-into-nine-ministers-over-microsoft-licensing-deal/

Fu, K., Cheng, D., Tu, Y., & Zhang, L. (2016). Credit card fraud detection using convolutional neural networks. In A. Hirose, S. Ozawa, K. Doya, K. Ikeda, M. Lee, & D. Liu (Eds.), *Neural Information Processing. ICONIP 2016. Lecture Notes in Computer Science* (Vol. 9949, pp. 758–759). https://doi.org/10.1007/978-3-319-46675-0_53

Galindo, J., & Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications. *Computational Economics,* 15(1–2), 107–143. https://doi.org/10.1023/A:1008699112516

Gallego, J., Martínez, J. D., Gallego, J., Rivero, G., & Martínez, J. D. (2018). Preventing rather than punishing: An early warning model of malfeasance in public procurement. Retrieved from http://repository.urosario.edu.co/bitstream/handle/10336/18525/dt222.pdf?sequence=4

Grace, E., Rai, A., Redmiles, E., & Ghani, R. (2016). Detecting fraud, corruption, and collusion in international development contracts: The design of a proof-of-concept automated system. *Proceedings of the IEEE International Conference* on Big Data, Big Data 2016, 1444–1453. https://doi.org/10.1109/BigData.2016.7840752

Greene, J. (2019, July 22). Microsoft to pay $26 million to settle probe into Hungarian kickback scheme. *The Washington Post.* Retrieved from https://www.washingtonpost.com/technology/2019/07/22/microsoft-pay-million-settle-probe-into-hungarian-kickback-scheme/

Hinshaw, D., & Greene, J. (2018, August 23). Microsoft hit with U.S. bribery probe over deals in Hungary. *The Wall Street Journal.* Retrieved from https://www.wsj.com/articles/microsoft-hit-with-u-s-bribery-probe-over-deals-in-hungary-1535055576

Hua, Z., Wang, Y., Xu, X., Zhang, B., & Liang, L. (2007). Predicting corporate financial distress based on integration of support vector machine and logistic regression. *Expert Systems with Applications*, 33(2), 434–440. https://doi.org/10.1016/j.eswa.2006.05.006

Leblanc, M., & Tibshirani, R. (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association*, 91(436), 1641–1650. https://doi.org/10.1080/01621459.1996.10476733

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. *Eighth IEEE International Confrence on Data Mining*. https://doi.org/10.1109/ICDM.2008.17

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2012). Isolation-Based Anomaly Detection. *ACM Transactions on Knowledge Discovery* from Data, 6(1). https://doi.org/10.1145/2133360.2133363

Matthews, C. M., & Ovide, S. (2013, August 21). Microsoft bribe probe reaches into Pakistan, Russia deals. *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/microsoft-bribe-probe-reaches-intopakistan-russia-deals-1377128173

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems,* 52(2), 464–473. https://doi.org/10.1016/j.dss.2011.10.007

Sill, J., Takacs, G., Mackey, L., & Lin, D. (1999). Feature-weighted linear stacking. *Foundations of Intelligent Systems* (ISMIS 1999), 592–600. https://doi.org/10.1007/BFb0095148

Smyth, P., & Wolpert, D. H. (1999). Linearly combining density estimators via stacking. *Machine Learning,* 36, 59–83. https://doi.org/10.1023/A:1007511322260

Wall, L. D. (2018). Some financial regulatory implications of artificial intelligence. *Journal of Economics and Business,* 100, 55–63. https://doi.org/10.1016/j.jeconbus.2018.05.003

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks,* 5, 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Zheng, Y. J., Zhou, X. H., Sheng, W. G., Xue, Y., & Chen, S. Y. (2018). Generative adversarial network based telecom fraud detection at the receiving bank. *Neural Networks,* 102, 78–86. https://doi.org/10.1016/j.neunet.2018.02.015

Zhou, Z. H. (2018). A brief introduction to weakly supervised learning. *National Science Review*, 5(1), 44–53. https://doi.org/10.1093/nsr/nwx106