Understanding the Differentiating Capabilities and Unique Features of Salesforce Einstein Discovery within the Machine Learning Space

Darvish Lee Shadravan
Distinguished Solution Engineer, Data Science & Machine Learning
dshadravan@salesforce.com

February 1, 2021

Abstract

The primary goal of this whitepaper is to elucidate the key features and capabilities of the Einstein Discovery machine learning (ML) product - specifically for readers who want to understand why it occupies a unique niche in the machine learning and predictive modeling marketplace. The intended audience for this document ranges broadly from business users and executives, to analysts, and of course to data scientists and machine learning professionals. This whitepaper is specifically intended for readers who want to understand why Einstein Discovery is different from other ML and analytics tools, and why it might matter to their organization.

Please note, while this whitepaper does cover many fundamental aspects of Einstein Discovery, it is not intended to be a comprehensive tutorial nor a general introductory training on Einstein Discovery; that task has been adequately addressed by many talented authors and bloggers. We include links to many of those resources in the appendix, and would encourage consumption of those sources if you are not already grounded in the basic functionality.

The approach we take in this whitepaper is to break down the differentiating capabilities of Einstein Discovery into three layers:

- Business Application Layer, covered in Section 2
- Data Science Layer, covered in Section 3
- Data Platform Layer, covered in Section 4

These layers are levels of logical separation and, in some cases, physical separation - similar to other common technology stacks (e.g. OSI network, full stack).
1 Introduction

Salesforce Einstein Discovery fills a unique niche in the crowded machine learning market by providing a compelling "power to the people" strategy. It embraces individuals with little or no statistical or data science training, while simultaneously bringing many benefits to trained data professionals. It provides world-class, supervised machine learning in a fully declarative environment, while also creating highly explainable and explorable data analysis targeted at a broad spectrum of potential users, which include:

- Business Analysts and Power Users
- Data Professionals, including Data Scientists and Machine Learning Practitioners
- Salesforce Users and Admins
- Business Managers and Executives
- Application Developers

While Einstein Discovery is capable of tackling a broad array of machine learning challenges, the majority of the real-world efforts performed with Einstein Discovery are focused on increasing transparency into business data problems, creating intuitive models, and operationalizing predictions and actionable insights into the hands of business users. Performing this level of analysis with the traditional ML pipeline approach is non-trivial, complex, time-consuming, and expensive. Adopters of Einstein Discovery are keenly interested in avoiding the lengthy development time, resource burdens, and hidden costs of building and deploying traditional custom predictive models. All models are initially
at least somewhat imperfect, but in many instances, getting a model that’s “close enough” into production and in front of business users is where Einstein Discovery delivers major return on investment. A clicks-not-code prototyping environment expedites analysis, streamlines model tuning, and realizes faster benefits in production. Einstein Discovery facilitates a rapid, iterative revise-and-redeploy approach that engages operational feedback (prediction accuracy) to ensure a solution roadmap with continuous improvement.

For data scientists, machine learning practitioners, and other statistical/data professionals, it is important to note that the role they will play in an Einstein Discovery solution deployment is more important than ever. Although their involvement may differ from their usual engagement, it’s still critically important that they bring their expertise to bear in these projects. An example from the author’s direct experience is a large, multi-thousand user Tableau CRM environment in which the goal was to create prediction scores that would be tightly integrated into opportunities and leads within their CRM application. While the IT admin team was able to facilitate much of the data gathering, configuration, analysis, deployment, etc., the model validation portion of the process was among the most critical keys to success. The data science team was heavily involved in feature engineering discussions, collaboration with the business experts on the use case, and validating the model performance. A critical contribution of the data science team involved an expert, in-depth comparison of the Einstein Discovery model with a more traditional coded solution. The performance, accuracy, and efficacy of the model turned out to be so close to the custom code solution that the difference was deemed irrelevant and not worth further scrutiny. In cases like these, the data science team is able to leverage Einstein Discovery as an enabling framework to operationalize and deploy models into large-scale production. That “last mile” implementation phase of data science solutions is often the hardest, most expensive, and (for data scientists) the least appealing part of the process. Einstein Discovery simplifies and streamlines the deployment portion of the data science pipeline, whether it’s for 10 users or 10,000.

Einstein Discovery began its life as a Silicon Valley startup named Beyondcore in 2004. The founders were inspired by the idea that “citizen data scientists” should be empowered to participate in the new wave of machine learning and AI for business use cases. The lineage of the firm is important context for this paper because of the incredible foresight the founders had with regard to building a world-class ML product that was aimed at the masses, accompanied by the unique patented technologies that sprang from this vision. In 2016, Beyondcore was purchased by Salesforce, primarily as a vehicle to add predictive and prescriptive machine learning to its already successful Wave Analytics Cloud (since re-branded as Tableau CRM). In 2018, Beyondcore officially became Einstein Discovery, and today rapid innovation continues with an entire team at Salesforce dedicated to its development and expansion. This team includes one
of the key founders and well-respected lead architect of Beyondcore, Mr. Griffin Chronis. One area of recent significant investment has been to deepen native integration with Tableau, the world’s leading data visualization company that Salesforce acquired in 2019. New integration capabilities allow Tableau users to get predictions and improvements in their Tableau data and visualizations by leveraging AI-powered Einstein Discovery models with clicks, not code.

2 Unique Einstein Discovery Capabilities at the Business Application Layer

Summary The business application layer as defined here is the interface with which the user will interact with the software - whether they are analyzing and exploring data, building and managing models, or consuming predictive and prescriptive insights in business applications. Einstein Discovery has unique strengths in the following key areas:

- Building Models
- Understanding Data
- Speed and Iteration
- Augmented Intelligence
- Workload Prioritization
- Automation
- Actionability

Model Building

Einstein Discovery has a unique interface for ease of model creation that makes no assumptions about data science expertise. A business analyst, for example, can create a "story" in Einstein Discovery simply by choosing a critical business metric they want to maximize or minimize. Once they've initiated the story building process, a graphical wizard interface walks them through the steps to choose which features to include in the model, add filters, review frequency tables, and so on.
At the completion of this simple graphical wizard, Einstein Discovery will (behind the scenes) execute a regression or tree-based analysis to create an ML model and also generate a collection of insights complete with visualizations and natural language-based explanations. The story is a key asset. In rich narrative form, it helps highlight statistical correlations between a business-relevant metric (an outcome variable, such as a KPI) and the explanatory variables that are potential influencers of that KPI.

The story metaphor as the user interface also provides explorable insights that allow the user to discover and diagnose patterns in the data with component-driver analysis, ask "what if" style questions, and, of course, review predictions and prescriptions in the model. All of that rich functionality is accomplished via a simple graphical wizard that requires no coding, nor any knowledge of machine learning mechanisms. This relatable framework allows the user to deeply understand the data with the full range of analytics capabilities - Descriptive, Diagnostic, Predictive, and Prescriptive.

**Data Understanding**

When Einstein Discovery is finished analyzing data, generating a story, and creating a model, the GUI presents the user with a series of insights. Each insight is represented with visualizations and natural language explanations. If the outcome (dependent) variable is numeric, a T-test statistical test is performed to determine which insights are most correlated with the outcome variable.
Likewise, if the outcome variable is a binary classification (yes/no, pass/fail, churn/not churn), then a Chi-squared test is performed. The insights in the resulting Einstein story are sorted according to statistical significance in the model. In other words, the first chart has the highest $R^2$ value relative to the outcome variable. The story insights also list the predictor variables with the strongest correlations relative to the outcome in the model. (Note that the $R^2$ values for all coefficients in the model are available in the model metrics tab).

Figure 3: Einstein Discovery Story Insights

Some of the descriptive bar charts are single variable representations, which are ranked purely by statistical significance. As the user scrolls further into the story, they will encounter charts where two variables have been combined in order to deepen their understanding of the data. For these two-field (bivariate) bar charts, Einstein Discovery uses a technique that calculates the adjusted $R^2$ of the coefficient for each combination of values. This powerful mechanism means that bar charts with two variables are sorted by deep interaction effects that could not have been understood statistically by looking only at univariate barcharts.

Einstein Discovery provides a predictive "what if" functionality by allowing the analyst to simulate prediction scores interactively. In addition to delivering a prediction score, the interface also shows top explanatory factors and ways to improve the predicted outcome. This capability allows the user to interactively simulate predictions for groups/categories within the model features.
Figure 4: Einstein Discovery Predictions

Communicating about the data, with visualizations and automatic explanations of how individual data elements are interrelated, has always been at the forefront of Einstein Discovery’s design philosophy. The following text is pulled directly from Beyondcore Patent 9,135,286 filed in 2015: "Variable value combinations that are predominant drivers of key observations are automatically determined from several competing variable value combinations. The identified variable value combinations can then be then used to predict future trends underlying the business intelligence data. In another embodiment, an observed outcome is decomposed into multiple contributing drivers and the impact of each of the contributing drivers can be analyzed and numerically quantified—as a static snapshot or as a time-varying evolution. Similarly, differences in observations between two groups can be decomposed into multiple contributing sub-groups for each of the groups and pairwise differences among sub-groups can be quantified and analyzed."

**Speed and Iteration** A common challenge for many organizations is that once built - a well-crafted model can be difficult to implement into production environments and integrate seamlessly with the operations they are designed to benefit. A key differentiator in the Einstein Discovery value proposition addresses the total quantity of human effort that’s required for the creation, deployment, and operational integration of a typical machine learning model. Consider the aggregate amount of time it takes for a typical data scientist to go from a raw dataset to a fully trained ML model that delivers descriptive, diagnostic, predictive, and prescriptive insights derived from the data. When looking at the spectrum of available ML tools, it is clear that there is significant
variance in the total time required for this series of events, ranging from minutes to weeks or months. Einstein Discovery minimizes the time for model creation, deployment, and tuning, thus increasing business value in cases where ”time to market” rapid delivery is critical.

**Augmented Intelligence**  Most ML products exist in a ”data vacuum”. Datasets are typically extracted, wrangled, and stored specifically for the purpose of model creation. Einstein Discovery is fully integrated with the Tableau CRM (formerly Einstein Analytics) data platform, which is intrinsically part of the larger Salesforce SaaS environment. Einstein Discovery enjoys extensive, native integration capabilities that go well beyond simply producing models. Therefore, model output can augment intelligence across a variety of mechanisms. Users can mash up descriptive, predictive, and prescriptive types of data assets within the same platform, all derived from the same datasets. This comprehensive framework allows admins and developers to rapidly create highly useful apps that deliver the full spectrum of data insights and embed them strategically in operational workflows.

![Figure 5: CRM Application Augmented with Tableau CRM Visualization and Einstein Discovery Predictions/Prescriptions](image)

**Prioritization**  Einstein Discovery insights and models help users prioritize and focus their efforts on the areas where they can most favorably affect business goals (e.g. which of my customers are most likely to churn, and what can I do about it?). Key business goals and metrics almost always boil down to minimizing or maximizing something, such as ”maximize cross-sell”, ”minimize customer attrition”, ”maximize lead conversion”, and so forth. The simple but powerful idea of maximizing or minimizing a key outcome variable (business goal) is fundamental to the way Einstein Discovery is designed. Its purpose-built to support rapid creation and iteration of machine learning models, and...
then provide predictive scores (with explanations) that allow business users to prioritize their workloads and make statistically supported business decisions. See this article for details and thoughts on prioritization: [https://www.linkedin.com/pulse/prioritization-hypothesis-darvish-shadravan/](https://www.linkedin.com/pulse/prioritization-hypothesis-darvish-shadravan/)

**Process Automation**  Automate robust business processes by embedding predictive intelligence in your process automation formulas. With the Einstein Discovery PREDICT function, flows and business decisions can be driven by automation logic based on predicted outcomes from Einstein Discovery. For example, in an approval process, a formula can determine whether a predicted outcome meets a threshold required for automatic approval. The PREDICT function is available when defining formulas for Next Best Action, validation rules, flows (screen, headless, and invocable), processes (in Process Builder), workflow rules, approval processes, predefined field values, field update actions, and default values.

![Figure 6: Predict Function Driving Business Process](image)

**Actionability**  For machine learning tools designed for professional ML practitioners, injection into the business workflow is often an afterthought. Their primary goals commonly end at model creation. However, a primary driver of ROI for machine learning investments is when organizations can leverage model output at run time to improve business outcomes.

Actionability in Einstein Discovery is a key differentiator from most ML tools. Actionable variables are variables that can be controlled, such as "place
customer on marketing journey” or “arrange annual policy review”. The analyst creating the model can designate one or more variables as “actionable”. Einstein Discovery applies prescriptive analytics on actionable variables to generate suggested actions that users can take at run time to improve the predicted outcome.

From Beyondcore patent number 9,098,810 - Recommending changes to variables of a data set to impact a desired outcome of the data set: "A method for recommending actions to affect an outcome of a process, the method comprising a computer system automatically performing the following: receiving from a user an identification of an outcome to be affected; processing a data set containing observations of the process, the observations expressed as values for a plurality of variables and for the outcome, wherein processing the data set determines behaviors for different variable combinations with respect to the outcome, the variable combinations defined by values for one or more of the variables; receiving an identification of one or more actionable variables from the plurality of variables; for pairs of a first variable combination and a second variable combination, wherein the first and second variable combinations are the same except that one or more of the actionable variables take first values in the first variable combination and take different second values in the second variable combination, predicting an impact of changing the actionable variables in the first variable combination from the first values to the second values by applying (a) the behavior of the second variable combination to (b) a population of the first variable combination; and recommending actions to change actionable variables based on the predicted impacts.”

3 Unique Einstein Discovery Capabilities at the Data Science Layer

Summary The data science layer as defined here includes the features and capabilities that allow Einstein Discovery to perform effective automated machine learning. This suite of capabilities includes augmenting, delegating, and offloading some of the more arduous and low-value tasks of data scientists. At the same time, these features retain the ability for deep inspection of models built by persons who are subject matter experts in the data but are not data scientists. Einstein Discovery has unique strengths in the following key areas:

- Model Training and Assessment
- Explainability
- Descriptive, Diagnostic, Predictive, and Prescriptive Analytics
- Algorithms
• Data Preparation and Cleansing
• Transparency and Trust
• Data Security and Bias Protection

Model Training and Assessment  In most data science projects, a common approach for model training and assessment is to randomly divide a dataset into three parts:

• The training set is used to fit the model
• The validation set is used to estimate the prediction error rate for model selection
• The test set helps provide an assessment of the final chosen model for which we are trying to generalize

To preclude the need to perform this assessment manually, Einstein Discovery uses the k-fold cross-validation technique on behalf of the user. Behind the scenes, when a user creates a story, Einstein Discovery uses part of the respective dataset to fit the model, and a different part to test it. Einstein Discovery randomly partitions the data into four equal size datasets. After model creation, Einstein Discovery provides model fit metrics for each of the four folds to quantitatively inform an individual reviewing model performance regarding over-fitting and other common modeling problems.

In Einstein Discovery, the H2O.ai K-fold function parameter is set at 4 (nfolds=4, so five models are built). The first four models (cross-validation models) are built on 80% of the training data, and a different 20% is held out for each of the four models. Then, the main model is built on 100% of the training data. The primary model contains the appropriate training and cross-validation metrics. All four of the cross-validation models contain training metrics (from the 80% training data) and validation metrics (from their 20% holdout data). For the main model, the four holdout predictions are combined into one prediction for the full training dataset. This “holdout prediction” is then scored, and the overall cross-validation metrics are computed.

Explainability  Historically, Machine Learning tools have been designed for data scientists and statisticians – not the business user who requires explainability and interpretability. Unlike other ML products, from its very beginning Einstein Discovery has always emphasized model interpretability as a core design principle. This key tenet has motivated the development team to use industry standard algorithms (described in “Algorithms” later in this section) to produce highly explainable and easily interpretable output. Hence, Einstein Discovery generates explorable visualizations and rich, natural-language narrative explanations in business-friendly terms.
With high model interpretability as a native capability, Einstein Discovery allows for the delegation of model creation (and resulting data analysis) to business analysts and other users who may not have data scientist skills. This is a key differentiator with Einstein Discovery. Within moments of generating a model on a dataset, the user sees a rich set of navigable insights derived from statistically-informed analysis and ML on a massive scale. Individuals who are subject matter experts in the data, but not necessarily data professionals, can rapidly create, iterate, interpret, and deploy predictive and prescriptive models.

In one of the five key original patents for Beyondcore (Patent number 9,129,226), the following statement was made regarding analysis of datasets by "inexpert" humans: "A combined computer/human approach is used to detect actionable insights in large data sets. Automated computer analysis used to identify patterns... These are presented to humans for feedback, where the humans may have little to no training in the statistical methods used to detect actionable insights."

**Descriptive/Diagnostic/Predictive/Prescriptive**  It is rare for a machine learning product to produce output that crosses the entire spectrum of major analytics functionality with explorable visualizations, natural language explanations, "what if" capabilities, a comprehensive analysis of the dataset, predictions, and prescriptive suggestions on how to improve predicted outcomes based on any chosen actionable variable. When reviewing tools that will be used across a variety of personas, including some without statistical skills, these capabilities represent a key driver of ROI.

**Algorithms**  In order to provide best practice statistical analysis that is optimized for interpretability and explainability, Einstein Discovery uses industry standard algorithms provided in H2O.ai, an algorithm library that is familiar to data scientists across industries. Einstein Discovery models use such modeling algorithms as:

- GLM (Generalized Linear Model) - either a piecewise linear model with ridge regression, or a piecewise logistic model with ridge regression
- XGBoost
- Gradient Boosting Machine (GBM)
- Random Forest
- A pilot feature in Spring ’21 is a typographical clustering algorithm that adds Fuzzy Matching capabilities to Einstein Discovery. For categorical data, differences in capitalization, plurals, abbreviations, and variations cause ambiguity about how to group the data. Fuzzy matching enables automatic matching for categorical values that should be in the same bucket (e.g. CIO and Cio). Einstein Discovery uses the Levenshtein Distance typographical clustering technique. This new algorithm will assist
in smoothing over spelling variations, resulting in smarter categorizations and better predictions.

For regression-based models, Einstein Discovery uses L2 Regularization to minimize over-fitting. Optimal lambda $\lambda$ values are automatically determined and applied on behalf of the user.

Figure 7: Ridge Regression Formula

\[
\sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2
\]

Figure 8: Ridge Regression (L2 Regularization) is depicted visually below. L2 seeks to improve on the Ordinary Least Squares (OLS) technique by imposing a penalty term to the cost function in order to prevent over-fitting and create a less complex model.
Piecewise linear models are single straight line linear graphs that many users are not familiar with. In Einstein Discovery, the piecewise algorithm fits straight lines for each decile of a continuous variable. This approach provides the benefits of regression simplicity and simultaneously adds the mathematical prowess to handle datasets that are not truly linear.

Figure 9: Typical two-part Piecewise Regression
When a user creates a story (and model) from a Tableau CRM dataset, Einstein Discovery automatically determines and presents viable outcome variables from all columns in the dataset. The target outcome can be a numeric continuous variable (for example, customer lifetime value) or a binary classification value (for example, churn/no churn or yes/no). The user selects whether to minimize or maximize this outcome, such as “Minimize propensity to churn (true)”. Based on the selected outcome, Einstein Discovery automatically selects the appropriate type of modeling algorithm to apply to any given model. GLM is the default algorithm for both linear and logistic models, but the user can alternately select any of the available algorithm options. In addition, ”Model Tournament” is an option which will test all available algorithm types for the current model. The winning model is selected with either $R^2$ for continuous variable outcomes or AUC for classification models. Although Einstein Discovery delivers model correlations and predictions utilizing GLM, GBM, XGBoost, or Random Forest, other algorithms are also utilized behind the scenes that assist with optimization of the primary model.
Data Preparation and Cleansing  To assist users who lack data science expertise, Einstein Discovery provides powerful mechanisms to detect and alert users about common problems that arise in datasets targeted for predictive modeling. This empowers users who are 'untrained' in statistical data methods to alleviate these issues and produce more accurate and reliable models. Data safeguards enable users to intervene and address the following kinds of issues:

- **Duplicates** - Multiple variables that provide essentially the same information; removing duplicate variables simplifies the model while maintaining accuracy
- **Outliers** - Observations with values that are unusually large or small, which affects calculated averages
- **Strong Predictors** - Variables that have a very strong correlation with the outcome variable; in certain cases, such variables should be excluded
- **Identical Values** - All observations in a variable belong to the same category, which contributes no value to the analysis
- **Dominant Values** - Most observations in a field are in the same category, which contributes little value to the model
- **Disparate Impact** - If recent story updates are affecting variables that the analyst has flagged as sensitive (having the potential for bias), Einstein Discovery may warn the user of disparate impact (in which variables that being treated unequally in the model)
• **Bucketing** - Einstein Discovery attempts to optimize binning of the data on behalf of the user

• **Cross Validation** - If cross-validation quantitative tests fail, Einstein Discovery alerts the user

Outliers (extreme values) are one of the most common problems in machine learning datasets. Because Einstein Discovery is sensitive to these outliers, the product automatically identifies data values that are greater than five times standard deviation, and recommends that the user remove them. The user can also reduce outlier influence by using transformations or converting the numeric variable to a categorical value with binning.

Furthermore, Einstein Discovery identifies strong or "obvious" predictors with an $R^2$ value greater than 0.3 so that the analyst can review them, determine whether they are valid features or problematic, and choose to exclude them from the model. Einstein Discovery helps users address other data issues, such as collinear values, excessive instances of null values, fields with only a single value, or fields with excessively high cardinality (many unique values).

For scenarios in which continuous value variables should be put into buckets, Einstein Discovery uses Kernel Density Estimation, an unsupervised learning
algorithm, to analyze the numeric columns in your story and suggests appropriate bucketing (ranges). Kernel Density Estimation is a well-known and widely-accepted method for visualising the distribution of data.

$$ \hat{f}(x; H) = n^{-1} \sum_{i=1}^{n} K_H(x - X_i) $$

* Where $K(x)$ is the kernel and $H$ is the bandwidth matrix.

Einstein Discovery approaches datasets with the notion that most business data is either implicitly or explicitly categorical, even when the data elements superficially appear to be numbers, dates, or text. This design approach allows Einstein Discovery (upon ingestion of the data) to divide all data into three distinct data types; categories, numbers, and dates. The user can explicitly choose the data type for each column, if they want, or they can let Einstein Discovery make assumptions about the data type based on the shape of the values.

By default, every unique value in a text field gets its own bin if it occurs with at least 1.5% frequency in the dataset. Any remaining values are assigned to the special “Other” category. Users have the option to filter out any values they want to exclude from analysis. They can also transfer any binned categories into “Other,” if they want. They can manually choose to allow values rarer than 1.5% to hold their own bins, but only if those values are in the top hundred observed values.

Variables with high cardinality variables (many unique values) can prove difficult to interpret and visualize. For this reason, Einstein Discovery ignores unique values above 100 in these variables or groups them into a reserve category. In a recent release, Einstein Discovery now allows the user to bypass this restriction and enable one high cardinality variable to a story (a maximum of 200 unique values). Einstein Discovery automatically raises an alert when it detects variables containing more than 100 unique values. The story creator can either let Einstein Discovery handle these variables as before or control how the variables are used.

Numerical columns are automatically divided into categories by decile (using a Kernel Density Estimation algorithm). This means that the bottom 10% of numbers get their own category, then the next lowest 10% get their own category, and so on. Users can adjust buckets (bins) to a fixed width, or they can crop the range as desired. If a numeric column has fewer than 10 different values, Einstein Discovery automatically converts the data type to text for a categorical representation. Date elements are intelligently bucketed into either trend or
periodic values, which an analyst can apply in a variety of typical temporal analyses. All of the aforementioned steps are simple options in the GUI. For the analyst, no coding or manual manipulation of the data is required.

Figure 14: Automatic Bucketing of a Continuous Numerical Variable

Einstein Discovery handles observations with empty or null values gracefully on behalf of the analyst. If the missing value is an outcome variable, Einstein Discovery omits that observation during analysis, does not factor it into averages,
and excludes it from the story insights. If the missing value is an explanatory variable used in a model, Einstein Discovery generates a warning rather than a prediction.

In Spring ‘21, Einstein Discovery introduces imputation (pilot), which allows the user to replace null values in observation with derived values. This ensures that observations are safely counted during analysis and that the model returns predictions instead of warnings. For example, suppose the annual policy amount column in an insurance agency dataset is missing many values. Imputation lets you replace those missing values with values derived from other data, such as the average policy amount by zip code.

**Templates** Einstein Discovery Templates offer an end-to-end workflow, enabling analysts to jump start an implementation and focus on customizations instead. The templates provide the quickest way to prepare, load, and analyze common business use cases with minimal clicks. In certain implementations, templates can replace the manual and complex process of model deployment. They can also automate the calculation and creation of dataset prediction fields.

**Figure 15: Einstein Discovery Templates**

---

**Transparency and Trust** Many traditional data science and ML products provide robust model metrics to assess model quality and performance. Most products assume, however, that the data professional using the tool knows
types of performance measurements they are interested in, and where to find them. Like many aspects of predictive modeling, Einstein Discovery employs traditional techniques and extends their capabilities in an aggregated and highly approachable format for the non-statistician.

Einstein Discovery’s Model Metrics section provides robust and relevant evaluation metrics to judge model quality and performance, including many of the most commonly used measures (Accuracy, $R^2$, Precision/Recall, AIC, RMSE, MSE, and so on). A data scientist assisting in model validation can rely on various qualitative metrics and statistical reports and charts, including prediction examinations, cumulative capture charts, cross-validation metrics, a visual Confusion Matrix, an ROC/AUC curve chart, and full coefficient listing. In addition, R code provides full transparency into the model’s transformation and scoring that Einstein Discovery created on behalf of the user.

Figure 16: Model Metrics - Model Evaluation

**Data Security and Bias Protection** Unlike other traditional data science products and tools, Einstein Discovery makes no assumptions about a user’s expertise in providing data privacy and security in an ML model. Einstein Discovery assists such users with multiple built-in mechanisms.

To ensure individual privacy within models, Einstein Discovery uses K-anonymity, which is a well-known and highly regarded statistical technique. Einstein Discovery provides a built-in mechanism to exclude data that might compromise or identify an individual within a dataset, such as a government ID number or phone number. Einstein Discovery implements K-anonymity in such a way that any data cluster of less than 35 rows is ignored, thereby ensuring that the model retains appropriate levels of anonymization and privacy.

Einstein Discovery also helps users create accountable AI models that detect and flag potentially biased variables. Einstein Discovery implements the Salesforce "Ethical AI" initiative, helping analysts prevent unintended statistical or ethical/societal bias. A user can designate a variable as a "sensitive variable" in order to selectively examine and optionally exclude specific data values from models. Einstein Discovery then notifies the user of problematic correlations.
To detect strong correlation (collinearity) with the sensitive field, Einstein Discovery calculates a Cramer’s V Score and raises an alert for anything above .5. Furthermore, Einstein Discovery proactively detects disparate impact in any variables that are being treated unequally in the model, and then notifies the model creator. This alert allows analysts to easily remove disparate impact bias from predictions, resulting in more ethical and accountable models.

**Bring Your Own Model (BYOM)** In pilot as of Spring ’21, Einstein Discovery enables data scientists to bring their existing models into the Einstein Discovery platform.

![Figure 18: Bring Your Own Model](from the Model Manager in Analytics Studio, a data scientist can upload and deploy externally created Python or TensorFlow models into Salesforce. This allows an organization to design, build, test, and tune models using their own data science tools and then operationalize them for Salesforce users. These external models can implement techniques that go beyond what Einstein Discovery
supports natively (for example, Deep Learning) as long as it is a TensorFlow model (Python can easily be converted to TensorFlow models). Many of the same benefits of the robust Tableau CRM platform are available for these external models, reducing deployment costs and implementation hurdles for teams tasked with deploying machine learning models into business operations at scale.

4 Unique Einstein Discovery Capabilities at the Platform Layer

Summary The data platform layer is a critical, but often under-appreciated, component of a machine learning solution. Einstein Discovery has the unique advantage of residing within the Salesforce Tableau CRM SaaS environment. Native embedding provides seamless integration with the big data storage, connectivity, enterprise-ready security, and robust data management capabilities inherent in the Salesforce cloud service.

Data Platform Data Platform Einstein Discovery is deeply integrated into the Tableau CRM Cloud. It uses Tableau CRM datasets to analyze data and store predictions, improvements, and top factors. Using the Tableau CRM framework, data engineers and admins can construct data integration solutions that interact with Einstein Discovery stories and models. Benefits include:

- A robust ELT (extract, load, transform) toolset and workflow framework
- Built-in connectors to most popular cloud data sources
- ML-based data transformations for sophisticated dataset data prep operations. These transformations utilize the Spark platform, which is familiar to most data professionals. Examples include sentiment analysis, clustering (pilot Spring '21) and Typographic fuzzy matching (pilot Summer '21).
- Visual tools for build workflows and data preparation
- Big data scalability. Tableau CRM datasets can manage up to 2 billion rows by default. Einstein Discovery can analyze up to 100 million of those rows per story.
- Public APIs, declarative elements, and programmatic access
- Scheduling tools that allow admins to specify when to run extract, load, and transform jobs
- Salesforce's secure, well known, and trusted cloud platform
- Sandbox environment with full datasets for testing and development, as well as the ability to package and promote tested models from sandbox to production environments
Model Management and Monitoring  Einstein Discovery provides model monitoring capabilities that allow administrators and data scientists to easily inspect, refresh, and manage models over time. This includes the following capabilities:

- Model monitoring tools provide real-time accuracy reporting
- Accuracy Analytics App provides Model Accuracy and Performance Dashboards
- Model alerts provide subscription-based notifications for model performance drift and out of bounds and missing values that exceed configured thresholds
- Automated model refreshes with a configurable schedule
- Side-by-side model comparison
- Configurable target thresholds for each model (for binary classification)
- Residual Plot Chart for logistic models
- Automatic snapshots of previous models
- Customizable evaluation order for predictions with multiple models (supports targeted predictions for segmented data)
- Model Cards to document and disclose important usage information about your predictions to others (pilot Spring ’21)

Figure 19: Model Management
Model Versioning  The individual(s) responsible for model management can track updates to models that result from changes in the data or improvements in story or model settings. The Einstein Discovery Model Manager displays a model’s version history so that the model owner can pinpoint exactly when it was updated and by whom, and whether it’s scheduled for an upcoming refresh job. For models that are not performing as expected, the user can easily revert to a previous model version with superior performance. Also, to investigate the underlying settings associated with a particular model version, the user can easily retrieve and examine the specific story version on which it’s based.

Deployment  For data science and machine learning projects, arguably the most challenging aspect of realizing sufficient return on investment is the effort and complexity required to operationalize models into production. The goal is to deliver models into business operations so that end users can benefit from consuming predictions, targeting prioritized workloads, and acting on suggested improvements in real time. When Beyondcore became Einstein Discovery, it was
integrated into Salesforce and engineered to natively interact with the existing Salesforce data platform (Tableau CRM). In addition to leveraging the enterprise features (actionability, security, etc.) mentioned previously, this integration critically allows for "point and click deployment" of a predictive model with real-time-scoring and performance monitoring. Using Einstein Discovery, we have seen customers build and deploy a model into production for thousands of users in a matter of days and weeks, well ahead of the typical lengthy ML project cycle.

Figure 21: One Click Model Deployment
**Einstein Prediction Services**  Although Einstein Discovery is built on the Salesforce Tableau CRM infrastructure, it is also capable of providing predictions to applications that are external to the Salesforce ecosystem. Einstein Discovery is designed to be agnostic regarding the data sources used to build a model. To support programmatic interaction with external applications, the Einstein Prediction Service API enables users to programmatically get, create, and manage predictions and associated models.

After an analyst creates and deploys a model with Einstein Discovery, a developer can use Einstein Prediction Service APIs to embed predictions into any website or application that is capable of consuming standard web service (REST) APIs. This enables data professionals to leverage the power of Einstein Discovery beyond Salesforce use cases. Data scientists, data engineers, and business intelligence professionals are able to integrate the scale, speed, and features of Einstein Discovery into their existing data assets with all the aforementioned benefits. To learn more, see [https://help.salesforce.com/articleView?id=bi_edd_prediction_service.htm&type=5](https://help.salesforce.com/articleView?id=bi_edd_prediction_service.htm&type=5) in Salesforce help.

**Einstein Predictions and Tableau** As the Tableau and Tableau CRM (formerly Einstein Analytics) products are now under the same roof, integration work is rapidly coming to fruition between Einstein Discovery and Tableau. As of the Spring ’21 release, an analyst can get predictions for Tableau data using
Einstein Discovery models that are deployed in Salesforce. To link to a model, an administrator can connect to Salesforce, generate a custom script for the model in Model Manager, create a calculated field in Tableau, and then paste the script into the Calculation Editor for that field. Then, when the script runs, it sends a prediction request with the required data to Salesforce. The Einstein Discovery model calculates a prediction and returns it to Tableau, where it appears in your worksheet at lightning speed.

Figure 23: Einstein Predictions Surfaced in Tableau

This ability to easily inject ML predictions directly into a Tableau visualization with no code will appeal to many analysts, users, and data science professionals who use Tableau daily. More native integration features with Tableau will be announced soon as part of the Tableau 2021.1 release. To learn more, see https://help.salesforce.com/articleView?id=release-notes.rn_bi_edd_tableau_tablecalc_script.htm&type=5&release=230
5 Resources and Further Reading

For reference, training, and context, we provide the links below as a starting point. We also encourage reviewing both the Gartner BI and the Machine Learning Magic Quadrants to understand the point of view of the largest industry analyst. Note that Salesforce is currently listed in the BI Quadrant because Einstein provides descriptive analytics as it’s most well known capability.

https://help.salesforce.com/articleView?id=sf.bi.edd_learn_more.htm&type=5


https://trailhead.salesforce.com/en/content/learn/trails/wave_analytics_einstein_discovery


https://www.forcetalks.com/blog/advantages-of-salesforce-einstein-discovery/


https://www.salesforceblogger.com/2019/10/30/take-your-ed-model-from-good-to-great/

https://www.linkedin.com/pulse/prioritization-hypothesis-darvish-shadravan/


6 Credits

My sincere appreciation to Eric Ivory-Chambers for his willingness to dive in and provide world-class editing of this paper.

A very special thank you to the entire Einstein Discovery Product Management team for their collaboration and support on the creation of this paper.