

ADVIZOR Modeling Workbook

This workbook guides an ADVIZOR Analyst/X user through the process of building predictive models.

This workbook contains three major sections:

1. **General Concepts:** This introduces the definition, components, and thought process of predictive models. It concludes with the eight step process we recommend for building models.
2. **Application:** The application section covers the modeling interface within Analyst/X.
3. **Demo Walkthrough:** Finally, we walk through the process using the “Customers” demo that comes with the software.

How to Read:

Supplementing the main body of text and scattered throughout the workbook are several types of textboxes.

Doug’s Helpful Tips – These are suggestions, thoughts, or comments from ADVIZOR CEO Doug.

Mateo’s Main Point – These are summaries of each section from ADVIZOR’s lead business consultant Mateo.

Linear Regression Case Study – The application of the section’s topic in a case study that predicts a target that is a numeric value (Linear Regression).

Classification Regression Case Study – The application of the section’s topic in a case study that predicts a target that classifies data into two categories (Classification).

Table of Contents

How to Read:	1
--------------------	---

General Concepts

What is a Predictive Model?	3
The Business Question	3
Linear Regression vs. Classification	4
Base Population	5
Target Field	6
Explanatory Factors	7
Thinking Through Factors	7
Model Training and Output	11
8 Step Process	12

Analyst X Application

Using the Predictive Modeling View	15
Creating a New Model	15
Current Model	15
Advanced Options	16
Exclude Rows from Prediction	16
Configuring a Model	16
Creating a New Model	16
Understanding the Model	18

Demo Walkthrough

Step 1. Define the business question in terms of data	20
Step 2. Select the target field and identify the subset of the data that is relevant to the question (the base population)	21
Step 3. Visually Explore and Form Hypotheses	21
Step 4. Select Explanatory Factors	23
Step 5. Train the model	24
Step 6. Examine the model and iterate	25
Step 7. Integrate output into the project	27
Step 8. Continually evaluate model performance	27

General Concepts

This section covers the basics of Predictive Modeling, starting with the definition of predictive models and how to approach predictive modeling with business questions in mind. The following sections cover the inputs: the relevant population, the target field, and the explanatory factors.

What is a Predictive Model?

A Predictive model is a mathematical and statistical description of patterns in a set of data. The model then applies to a new set of data to make predictions. Here are some examples:

- Using data on past accidents to calculate the risk of accidents for policy holders in car insurance.
- Using demographical and behavioral data to analyze which appeals will yield the most gifts in fundraising.
- Using internal and external economic indicators, to predict the price of shares in a company.
- Using previous consumption information to estimate the amount a customer will spend while visiting a store.
- Using historical trends to predict the demand for certain machines in manufacturing in the coming months.

In each of these cases, a model can describe the relationships between the data and the outcome. Models are built on a single table that contains all of the relevant information.

The Business Question

A business need should drive every model. The first step is to translate that need into a question that can be modeled. A good question has a specific set of data, a knowable outcome, and an actionable result in mind. Here are some example questions:

- Which customer segments are likely to respond to an advertising campaign?
- Which prospects are likely to make large donations?
- How effective is a new medicine at treating a certain disease?
- How many calls will a customer support center receive on a given day?

With each of these questions, there is a particular population or subset of data to be examined. There is also an outcome to predict that is present or can be represented within the data. This outcome has a clear potential actionable item as the answer. Without a knowable outcome to predict, we cannot train the model to identify the outcome. Finally, without a potential clear actionable result, the model becomes a “fun facts about the data” exercise. Rather than asking broad questions, it is better to take existing business needs and formulate them into actionable questions.

Mateo's Main Point: Predictive models should be driven by business needs and you should know what actions you will take based on the results.

Linear Regression vs. Classification

ADVIZOR Analyst/X uses two modeling algorithms: Linear Regression and Classification. (The particular technique used for Classification is “Logistic Regression”.) The two types of regression are used for two different types of questions.

- **Linear Regression** predicts a numerical value. Examples include predicting the temperatures in the coming week, pricing of various stocks and bonds, the expected unemployment rate in the coming months, or the agricultural output of a farm given the growing season.
- **Classification or Logistic Regression** looks at a binary outcome (success or failure, win or lose, will purchase a product or not, yes or no), in order to predict the likelihood of a future occurrence. Classification gets its name from its common use: classifying the population into two groups. One group is associated with a preferred outcome. In this case we want to see how likely is someone to become a part of the preferred group in the future.

Mateo’s Main Point: Linear predicts “How much?”

Classification predicts “Will?” or “How Likely?”

Case Studies

Linear Regression Case Study: Pool Supply Call Center

A Pool Supply Company has a customer service call center. The company would like to be able to forecast the number of calls on a given day. This will allow them to optimize the number of staff needed to answer the customers in a timely manner. The business question can be phrased as “For a given day, how many calls do we expect?” This has a specific set of data (the historical call center data), a knowable outcome (the number of calls), and a clear actionable result (adjusting the number of staff on the call center). This is a Linear Regression model because the outcomes (the number of calls) can take a range of numeric values.

Classification Regression Case Study: Prospect Identification

A university would like to find million dollar donors and has many alumni in its database. The business question can be phrased as “Of the alumni, who are the most likely to donate a million dollars?” This is a classification model because the outcomes can be one of two things: either an individual will give a million dollars or more, or they will not.

NOTE: Technically this could also be run as a linear regression model, but trying to forecast a skewed distribution like this is hard – hence the preference to first forecast it as a “set membership” classification. The forecast will be skewed because the vast majority of alumni will never be able to give a million dollars, yet there will be a handful who will give amounts much larger than a million dollars. Distributions like this are hard to forecast, and we would recommend doing this as a second step, if at all, and comparing the results with the classification model.

Base Population

The base population consists of the people, items, or occurrences that are of interest in our model. The base population contains existing data used to insight into what is associated with our desired outcome. For a model predicting the temperature on given day, the base population is the historical data for previous days. For a model on whether a customer likely to respond to an advertising campaign, the base population would be all the customers who had previously been exposed to the campaign, some of whom responded and some who did not.

Mateo's Main Point: Use a subset of your data that looks like the population whose behavior you want to predict.

In the data, the base population is represented by the rows of the table the model is to be built on.

The base population should made up of entities with similar experiences and are capable of the same behavior. For example, if we have a list of donors but it includes individuals and corporations, the experiences and influences of these will be very different –i.e. a corporation will not have a gender, cannot attend events or respond to emails. So, we might want to build a model only on individual donors. Instead of looking at all the data in a table, we might look at a smaller subset of rows. Additionally, all entities in the base population should have the capability of exhibiting the outcome. For example, if want to model voting in an election, we wouldn't want to include those under 18 in the base population.

The size of the base population must be large enough for the modeling algorithms to make statistically significant conclusions and extrapolate future predictions from past results. A safe rule of thumb is to have at least 30 rows per explanatory factor.

Case Studies

Linear Regression Case Study: Pool Supply Call Center

To build a model on the number of calls we expect in a day, we need a table with historical information at the day level. That is, a table which is constituted of 1 row per day, with columns for characteristics about that day. The base population is each of the historical days. In this scenario, we have 5 years of call center data. If we have multiple types of calls, we can build separate models for each type of call.

Classification Regression Case Study: Prospect Identification

To build a model on which alumni is most likely to donate, we will need data at the alumni level. Though the university has data on many affiliated entities, our base population will just be alumni. Non-alumni don't have class years or student activities and don't attend reunions. Thus, if we want to use those factors in our modeling, we will need to limit it to just alumni or else those factors will be biased against non-alumni since they are not able to have these experiences.

Target Field

The target field is a column in the data that describes the outcome to our question. For linear regression, the target field contains the values we would like to predict. For a classification model, you need a field with values of “0” or “1”. You might have such a field in your data, or you may create one using an expression.

The target field is also known as the “dependent” or “response” variable. It is the variable that is “dependent on” or “responds” to changes in the explanatory factors. For classification, it is convenient to use the term **Target Population** to describe the subset of data that exhibits the target behavior.

Mateo’s Main Point: The target field represents the outcome to predict.

Doug’s Helpful Tips: You can also build a classification model on the “selection state” in ADVIZOR Analyst/X, the set of rows that have been graphically selected. You can conveniently do graphical selection to identify the interesting population, and then build a model to predict members of that population.

For a classification model, there must be an appropriate ratio between the target (value “1”) and base population. The target to base ratio should be at least 1 to 200. If it is too small then you will need to reduce the base population. For example, if forecasting major giving, then you can cut the base to people who are rated over \$50K or something similar. You can still score everybody even if they are not in the base. In addition, there should be at least 30 rows relating to the favorable or positive outcome that we are trying to predict.

Case Studies

Linear Regression Case Study: Pool Supply Call Center

For our call center, we will try to predict the number of calls received on each day. Thus, in our table, we need one column that contains the number of calls received in previous days.

Classification Regression Case Study: Prospect Identification

The target population should be alums who have given a million dollars or more to the university in the past. In the case of the university, there are 500 alumni who have given over a million dollars to the university out of a pool of 50000 alumni. The ratio for this is within acceptable limits to ensure significant extrapolation from the target population.

Explanatory Factors

The explanatory factors, also called independent variables, are things that we consider to have a potential influence on the target outcome. Often, they come from hypotheses we have about the data and its relationship to the target. Each explanatory factor is a column in the table containing the base population. Explanatory factors can be a quantitative measurement or a categorical characteristic.

Doug's Helpful Tips: A robust model will typically run with 10 to 15 explanatory factors. We sometimes see teams trying to work with far more – 190 in one case. That's almost always a bad idea because the model will rarely have enough data for that and the result will suffer greatly. The technical term is "overfitting".

Examples of quantitative explanatory factors include the income of a household, the amount of a treatment given, or an individual's height and weight. Examples of categorical factors include marital status, the type of advertising campaign (mail, email, telephone...), or the state of residence. Data for explanatory factors can also be brought in from multiple tables into the main modeling table.

One distinction to make is between "correlation" and "causation". The model results show correlations between

the explanatory factors and the target field, and that may not always imply causation. "Correlation" means that two values vary together consistently. This correlation may be because the target is "caused" by the explanatory factor but this is not necessarily true. See *Independent vs. Dependent*, and *Confounding and Lurking Factors* in the next section. We need to think through the explanatory factor and how they relate to the target in order to infer causation in the relationship.

Mateo's Main Point: Explanatory factors are other data used to explain and predict the model target. Explanatory factors must be in the same table as the target field.

Thinking Through Factors

It is important to think through and discuss the explanatory factors and identify potential sources of error and bias. Here are some additional considerations:

- **Names or Keys** – Fields that uniquely identify an entity, such as a name or key IDs, are not useful for modeling. These are arbitrary and you do not want to include them in models.
- **Independent vs Dependent** – Explanatory factors should not depend on, or be caused by, the target field. For example, if the target population is those who made a large donation, we do not want to include attendance at a donor recognition event as a potential factor. Such a field would result in a model that perfectly predicts the target, but doesn't tell us any information we don't know. In general, a perfect model is a bad model.
- **Binning** – Binning refers to grouping values in the explanatory factor. Binning applies to both quantitative and categorical factors. For quantitative factors, binning reduces the impact of noise, outliers, and highly skewed data. Quantitative factors can also be binned into groups that we reason a priori to have similar impacts on the target field. A common example of this is grouping age into certain demographic segments, i.e. under 18, 18-35, and so on. If we make a

chart on disposable income vs age, the data points would not create a straight line that increases from young to old. Binning the ages helps the model to capture the non-linear impact. For categorical factors, binning assists in analyzing data with low counts in each category. For example, a doctor's specialty can be very specific and certain smaller specialties might have low numbers of doctors. Binning the specialties into more general groups can ensure that there are enough observations in each group to have statistically significant results. Binning categorical factors also assists in reducing the number of possible values for large cardinality factors. Binning groups these miscellaneous bits of data into its own category. This increases the speed of modeling and might pick up on an otherwise unknown trend if each of the possible miscellaneous categories were considered on their own.



- **Dates** – Dates are not useful in a model since the relationship to a specific date in the past is not useful. It is better to look at certain characteristics of that date, such as day of the week or

month. Consider whether there are cyclical trends by month or year. You can also compare two dates. Examples include the time since purchase, the time between multiple contacts, or the number of days from today. ADVIZOR does not allow dates to be used at explanatory fields. Instead, you should create additional fields with the Date Parser based on the date.

- **Locations** – Location and address values at the street level are too granular for modeling purposes. Rather, group the data at the city level or higher. The numeric value of a postal code is not useful, but other data about the postal codes might be more useful. Household income level, age distribution, or other demographic information is much more useful for modeling. You can also use location to calculate the distance from a specific point, such as store location or major city.
- **Bias in Data Collection** – It is important to understand the source of the data for the explanatory factors and the biases in data collection it might bring. Often, we won't have complete data for a given explanatory factor; some of the rows might have missing data. ADVIZOR's modeling can work with missing data. However, we should avoid a factor for which there is a systematic bias in collection. For example, we should avoid a factor that only a subset of the population has. As mentioned previously, it would be better to adjust the base population. Another example is survey data taken from an event will skew information towards those who attend events. Another source of bias is systematic measurement bias. This is bias that consistently offsets values by a certain amount. For example, a scale for weighing incoming shipments might always read off values by several pounds.
- **Confounding and Lurking Factors** – A confounding factor (Z) is one that is correlated with both the target field (Y) as well as another explanatory factor (X). The target field (Y) and the other explanatory factor (X) will appear to be correlated, when it is really the confounding factor (Z) that is correlated. A lurking factor is a factor that is not taken into account in the model. It has similar effects as a confounding factor. For example, if we look at incidents of drownings at the beach and ice cream sales at the beach from month to month, we would notice that the two are correlated. We might then infer that ice cream sales cause drownings. In reality there is a third factor, the season or temperature. More people are at the beach during summer, causing more ice cream sales and likelihood of drowning incidents. Not accounting for confounding or lurking factors can cause spurious conclusions about the data.
- **Multicollinearity** – Multicollinearity occurs when several factors are highly correlated with each other. This can cause overfitting, inaccurate assessments of the impact of each of the correlated factors, and high sensitivity to the correlated factors. The overall prediction will still be accurate, but be weaker with regards to each individual factor. There are two options for dealing with multicollinearity. First, we can re-examine and remove one of the factors. After all, if two factors are highly correlated, we don't lose information taking one of them out. Alternatively, we can use a larger sample size or obtain more data. The factors might have just happened to be collinear for that data set only and larger data sets might reduce the correlation between them. Finally, we can choose to ignore the multicollinearity, keeping in mind the potential consequences above.

- **Outliers** – Outliers are data points that are far from the rest of the data and have extreme values. Outliers can come from bad data, unique one-off cases, or simply due to random chance. For example, a data entry error may cause an extra 0 to be typed in giving a value 10 times what is expected. There are several ways to adjust for outliers, though the matter is subjective. The first is to simply exclude the data point. If the reason for the outlier is clearly a data entry error, then excluding it makes sense. However, because some amount of outliers can be expected with large data sets, most would shy away from outright excluding. Alternatively, one can Winsor the data. This means taking the outlier value and replacing it with a less extreme value and trimming the range. For example, if most data points lie between 70 and 110, we would set any value beyond 110 to 110. Finally, one can keep the outlier. Some outliers are expected for large data sets so removing them is unnecessary.
- **Overfitting** – Overfitting is when a model describes the data too specifically. Rather than describing large overall trends, the model describes every specific instance. Although it is extremely accurate for the existing data, it has poor predictive capabilities. Overfitting occurs when there are too many factors for the size of the target field.
- **External Data** – Explanatory factors can come from other tables beside the main table used for modeling. You can bring data from external sources into the main modeling table. Often data will need to be aggregated in some way to get it into the same level of detail as the main modeling table. For example, data from a sales transaction table should be aggregated to the customer level and brought into the customers table. Sales data can be aggregated many ways, including counting the number of transactions per customer, the largest transaction per customer, or whether a customer has bought a certain product.

Case Studies

Linear Regression Case Study: Pool Supply Call Center

To estimate the volume of calls in the call center on a given day, possible explanatory factors include factors related to business operations and factors related to the day. Factors related to other business operations include sales history, the introduction of a new product, or whether there is an advertising campaign underway. The day of the week or the month can have an impact. More calls might come in during the summer months as pool usage increases or on Mondays after problem is identified during the weekend.

Classification Regression Case Study: Prospect Identification

Possible explanatory factors for alumni likelihood to give include demographic information about the alumni, engagement in recent years, and information about their time at the university. Demographic factors include their age, their distance from the university, the median income of their ZIP code, or whether they are married to another alum. Behavioral and activity could include attendance at reunions, the number of other events attended, or whether they buy season tickets to athletics events. Information about their time at the university could include their degree and major, what sports they played, and what student organizations they were a part of.

Model Training and Output

Once the modeling algorithms run against the training data set, the model is created. Training a model finds which explanatory factors correlate with the target field. It finds the magnitude and direction of the effect, whether positive or negatively correlated. These factors combine together into an equation that calculates the predicted target field value. For linear regression, this is the actual predicted value of the target field. For classification, the model calculates a score that corresponds to the likelihood of an individual belonging to the target population. This model equation can then be applied to new data to predict future outcomes or to data outside of the initial base population.

Not all explanatory factors will be relevant to the model. Each explanatory factor is evaluated to determine whether its variation is statistically significant to the variation in the target field. There must be a large enough correlation between the change in the factor and change in the target. The statistical threshold for

Mateo's Main Point: The model shows which explanatory fields are correlated with the target field, how they are correlated, and creates an equation that uses these to predict a target outcome.

significance is known as the P-Value. It is the percent chance that the result that we have in the data is due to random fluctuations and not due to an underlying trend. Common values include .1, .05, and .01, with smaller values indicating a smaller percent chance of random variation.

Each relevant explanatory factor is included in the model equation with a coefficient and possible transformation. The coefficient represents the change in the target field corresponding to one unit of change in the explanatory factor. Positive and negative coefficients indicate positive and negative correlation respectively. Transformations account for non-linear relationships such as factors that have diminishing marginal returns. Examples include taking the square root or the logarithm of a factor.

A quality metric accompanies the model and indicates how robust the model is. Linear regression models use the R^2 value which measures the proportion of variance of the target field explained by the model. Typical values for R^2 between .7 and .9, though when trying to predict human behavior, even .5 is acceptable. Logistic regression models (aka classification models) use % Concordance which is proportional to the number of predicted values that match the actual value (which for logistic regression is either 1 or 0). The % Concordance typically runs between .6 and .95. For either score if the model quality is too high, it is indicative of overfitting. You may be attempting to use too many factors or using a factor that is identical or a proxy for the target. A low model quality indicates the chosen factors do not predict the target value accurately. Additional factors are needed that can better explain the target.

It is important to examine the model output and ensure that the results make sense. Go through each of the significant factors and check the results against common sense. Try experimenting with other factors, iterate through the process, and continue to refine the model. Create some charts to display the fields, calculate different transformations of the factors, and discuss the results with the team.

Case Studies

Linear Regression Case Study: Pool Supply Call Center

The model for the pool supply call center revealed that the call volume increases on Saturday through Monday, increases in the summer months, and with increased sales in the previous month.

Classification Regression Case Study: Prospect Identification

The model for the prospect identification found that alumni who have given in the last five years, who played sports during their time at the university, and who are in their 50s are the most likely to give to the new athletics center. Additionally, the probability increases with the number of events attended and the number of years they were season ticket holders.

8 Step Process

The eight step process we recommend when building models is:

1. Define the business question in terms of data
2. Select a target field and identify the subset of the data that is relevant to the questions ("Base" Population)
3. Visually explore and form hypotheses
4. Select explanatory fields
5. Train the model
6. Examine the model and iterate:
 - a. Are the correlation coefficients ((% concordance or R^2) in an appropriate range
 - b. Discuss whether the explanatory fields that scored high make "sense". Are they:
 - i. Truly independent, or could they be another manifestation of the target (e.g., attending the President's Dinner scored high, but only people who donate a lot of money are invited; etc.)
 - ii. Causal or just coincidental (talk through WHY the factor should have influence)
 - iii. Driven by and perhaps a subset some other factor (for example, county comes up high and so does city; then try one without the other and then reverse and see what happens)
 - iv. Etc.
 - c. If a factor scores really high, trying splitting into sub components:
 - i. For example, "# Activity Participations in the Last 3 Years" scores really high ...
 - ii. ... then run a second model against just the Activities Table and examine the range of Activities' influence – for example, some activities may be high (e.g., going to a reunion), some may be medium (e.g., going to a dinner), some may actually be negative (e.g., interviewing high school kids, many of whom don't get in)
 - iii. If this is the case, then maybe create "# High Activity Participations Last 3 Years", "# Neutral Activity Participations Last 3 Years", and "# Negative Activity Participations Last 3 Years"

- iv. And then rerun the main model with these 3 factors, vs. just one composite factor, and re-assess the outcome
- v. If the original factor had a low overall score, then splitting like this would not be necessary, and would have minimal impact on the overall model
- d. Re-run the model multiple times and discuss again
- 7. Integrate the output into the core reporting / data discovery systems or tools for general use
- 8. Continually evaluate model performance

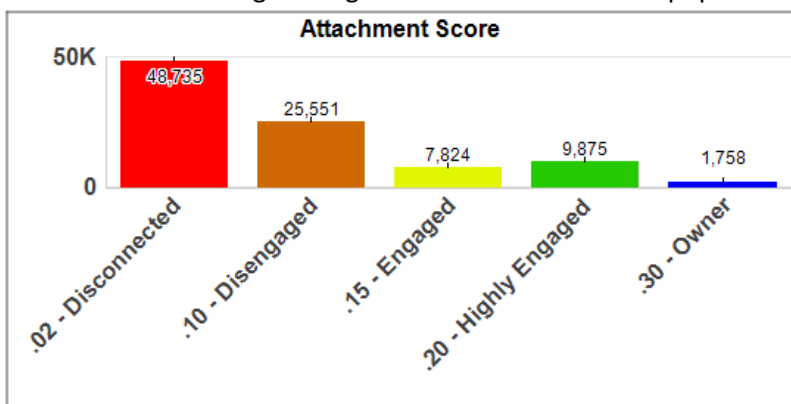
The model building process should be a collaborative effort with discussions at each step among the team. Discuss the main business question the model will answer. Brainstorm and hypothesize possible explanatory factors. The team should consist of both individuals who know the data and those who will use the end result. The team should focus on answering questions and start simple.

It is better to start with the data that you have and then iterate and evolve the model. A simple model completed now with perhaps just 5, 10, or 15 explanatory factors is better than a complex model that takes forever and perhaps could even overfit the data. We have seen examples of teams trying to work with 190 factor fundraising models, when there really are only a dozen or so things that matter, and not enough data to discern between many more factors than that – so the 15 factor model will actually outperform the 190 factor model and is MUCH easier to build and run.

Add new factors as they become available and evaluate the model performance in the months after the model is built. Embed the model in your data discovery or reporting system so that it updates.

Clearly Present the Model to the Team.

Remember that most not statistical team members will probably not relate well to modeling coefficients and numerical scores. So, we recommend binning the scores into 5 or so bins with very user-friendly names. In fundraising we might show attachment for a population in 5 bars, for example:



This format enables end-users to have a discussion about their highly engaged prospects without getting into a debate about why one is a .72, and another a .64 – which is often what happens if the raw scores are presented.

This being said, at some point in the discussion it is relevant to show enough detail so the team can understand where the current score came from, and what could perhaps be done to change it. So, in the example above we would also recommend having a page which has the details on each entity in the different groups, with the score and also the factors that make up the score. For example, here's a list of 11 highly engage prospects:

Attachment Score Details											
ID	Name	Rating	Total Lifetime Commit ▼	Attachment Score ▼	Attachment Group	#CommitteesL10Y	#GiftsLSY	#GiftsL6-10Y	#Reunions	#Sports	#Student Activities
26179	Matejko, Van	0 - NA	\$3,495	0.20	20 - Highly Engaged	1	5	5	0	0	4
188214	Viviers, Calandra	08 - \$50K - \$99K	\$13,540	0.20	20 - Highly Engaged	0	5	5	20	0	1
187903	Thongthai, Zachary	09 - \$25K to \$49K	\$4,440	0.20	20 - Highly Engaged	0	5	5	20	0	1
188962	Lawrie, Oscar	07 - \$100K - \$249K	\$4,100	0.20	20 - Highly Engaged	0	5	5	20	0	1
188300	Foster, Giuseppe	09 - \$25K to \$49K	\$3,858	0.20	20 - Highly Engaged	0	5	5	20	0	1
191888	Sorensen, Nigel	07 - \$100K - \$249K	\$1,000	0.20	20 - Highly Engaged	0	5	5	20	0	1
44720	Zhang, Deshaun	09 - \$25K to \$49K	\$1,245	0.20	20 - Highly Engaged	0	5	5	11	2	4
34596	Feuerstein, Aurelio	0 - NA	\$16,900	0.20	20 - Highly Engaged	0	5	5	13	1	3
29327	Sarma, Bernard	09 - \$25K to \$49K	\$3,923	0.20	20 - Highly Engaged	0	5	5	14	0	3
28706	Battistella, Graham	07 - \$100K - \$249K	\$1,403	0.20	20 - Highly Engaged	0	5	5	14	0	3
41429	Egan, Barney	08 - \$50K - \$99K	\$16,510	0.20	20 - Highly Engaged	0	5	5	12	1	3

This list provides the details on why the group is highly engaged in an easy to understand format. In this case all have made donations in all of the last 5 years, and also the 5 years before that, and all have been to a lot of reunions and played some sports and activities. Why they are highly engaged is pretty clear. But what also stands out is on one has been on a volunteer committee, and since the factors are arranged from left to right in order of priority there is a clear opportunity to increase the attachment scores of this group be getting them on the rightly volunteer committee.

Conclusion

Building an effective predictive model should be as much a discussion as it is a model building exercise. The most effective and most used models that we see are developed through an inclusive process that includes:

1. Discussing and agreeing on the key Business Questions to be answered
2. Choosing the Model Type that best addresses the business questions – typically either linear regression or classification
3. Discussing and defining the best Target, Base Population, and set of Explanatory Factors
4. Prepping the data for the Explanatory Factors
 - a. Where is the data? Can you get more?
 - b. Names or keys
 - c. Independent or dependent
 - d. Binning
 - e. Dates
 - f. Etc.
5. Training the model
6. Discussing the output with end-users; content the math to reality
7. Iterate the model, add / delete factors, re-run it, discuss the output again
8. Present the model to the user community in a friendly manner.

Analyst X Application

Using the Predictive Modeling View

The Predictive Analytics control panel lets you create models, edit existing model, or apply models. If the Predictive Model pane is not visible, click on the Model task bar and select the Model Data button

ADVIZOR Solutions - Mutual Funds



Creating a New Model

To add a new model: Click the “New Model ...” button. The model properties dialog displays. If a new model was trained, when processing is completed the model is displayed in two pages in the Analyst. The model that you’ve created must be evaluated for adequacy before it is used. The metric displayed depends on the type of model: A “Coefficient of Determination” is shown for regression models. This number, often called “ R^2 ”, give the amount of variability in the target that is captured by the model. For classification models, an “Adjusted Count” metric is shown. This is the number of predictions that match the actual values.

Current Model

The Current Model Grouping shows information about the current model. The “Models” combo box shows the name of the current model; this may be used to change to another model if you have more than one built. Operations on the current model are:

- **Edit an Existing Model** – Click the “Edit...” button to display the Model Properties dialog box. Make changes as required to the information displayed in this dialog

New Model ...

?

Current Model

Models: 3 Year Perf Model

Models

Edit ...

Delete

View

Copy

Cancel

Target Field: 3-Year Performance

Accesses (K): 1,152,378

R²: 78.4 %

Time: 00:00:00.27

Predicted Field: Predict_3 Year Perf Model

Model Type: Ordinary Regression

Training Set: 1754 of 1754

Change Models

Quality

Advanced Options

☒ Run equations with project.

Name bins ...

Exclude Rows from Prediction

Excluded rows during training:

☐ Do not predict rows from model where:

Limit Rows Predicted

Apply

box. Click “Train” to implement your changes, ADVIZOR builds the model. When the processing is completed, the model is displayed.

- **Deleting a Model** – Click the Delete Model button. A message, Delete model <Model Name>, is displayed. Click. The model is deleted.
- **Rebuild the Model** – Use the “Rebuild” button to re-train the current model with the existing target and explanatory fields.
- **View the Model**- If the model description pages are not visible, you may show them with the “View” button.
- **Run the Model** – Creates the predicted field values using the current model.

Advanced Options

Advanced Options are:

1. **Run equations with project:** If selected, the model equation is recorded with the project and will be run whenever new data is loaded with the project.
2. **Name bins ...** : You may name the binned scores that are produced for classification models to use other than the default bin numbers. Use this to give the group descriptive names.

Exclude Rows from Prediction

By default the model is applied to all rows in the training table, even those that were excluded during training. If you wish to constrain the model to a subset, you may enter an expression that identifies the rows that should not be included. These rows will be given a missing value for the predicted field. The top box describes the rows that were excluded during model training. To specify a subset to not model, check the “Do not predict rows from model where” check box. This will copy the excluded row expression from the first box to the second box. You may use this expression as is or edit the expression in the second box. Click the “Apply” button when the expression is what you want it and then click “Apply”.

Configuring a Model

The Model Properties dialog is used to configure the model in terms of its target field and explanatory fields. You reach it through the Predictive Modeling view.

Creating a New Model

Configure a model like this:

1. **Model Name:** Enter a unique name that identifies this new model in the "Model Name" field, or use the default name based on the table and target field.
2. **Data Table:** From the Data Table pulldown list, select a table that will be used to generate the predictive model.
3. **Target Field:** Select the Target Field from the pulldown list. This field answers the business question for the data being analyzed. Target fields may have either continuous values (e.g., real or integer numbers) or binary values (e.g., “0” or “1”). A target field cannot be a field with

multiple string values. You can create a field from the current selected field to use as the target, using the next button.

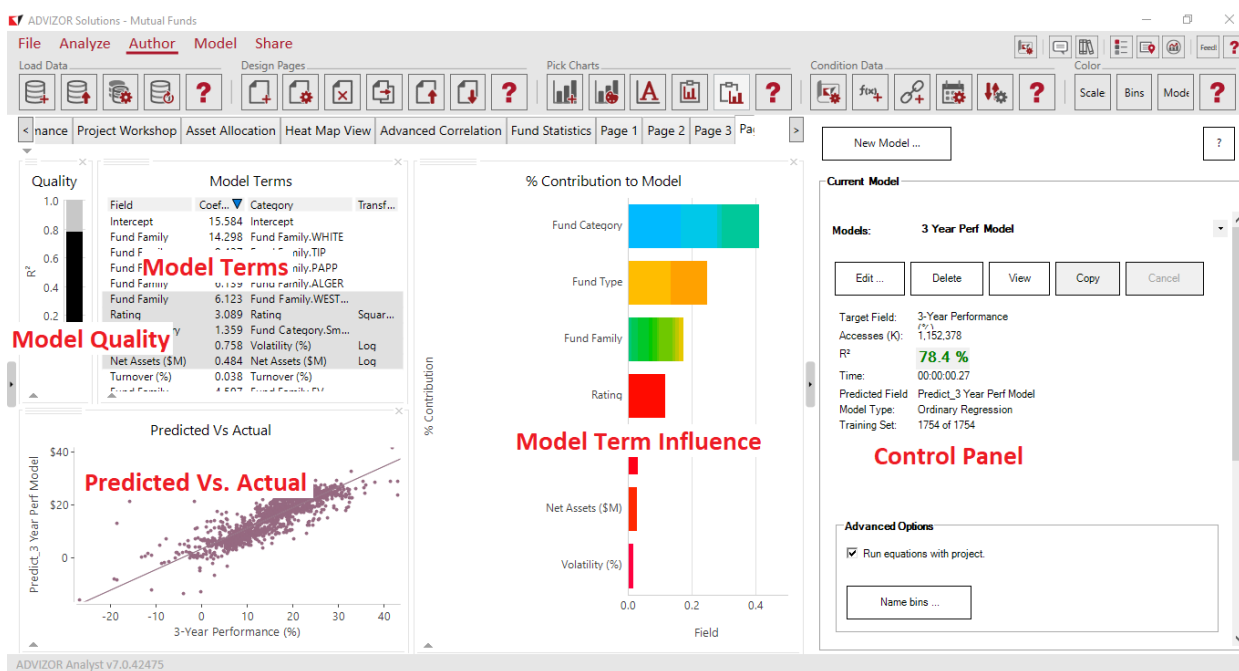
4. **Target from Selected ...**: Use this button to create a new field that contains the current selection state for use as a target.
5. **Explanatory Fields**: Explanatory fields are fields of the selected Table that will be analyzed to find those correlated with the target. Select one or more explanatory fields by clicking the check box next to the applicable field. Use the **All** or **None** buttons to turn on/off all fields. Not all fields can be explanatory fields!
 - The target field may not be an explanatory field.
 - Add data from other tables using the "copy" operation or "link()" expression in the Expression Builder.
 - Omit fields that are keys, where each row has a unique value.
 - Omit text fields containing arbitrary text that is mostly unique to each row, such as comments fields or addresses.
 - Date fields are omitted by default. You can translate them into an integer representing the elapsed time from a point in the past. You may want to calculate your own coding for dates, such as days from the present or days from a date in the past using the Expression Builder. You may also want to add additional fields that describe a date in terms of cycles (e.g., day of week, day of month, ...) to look for periodic patterns. These fields can be created with the Date Parser.
 - Zip Codes may require special modeling
 - String fields with large numbers of categories are omitted by default. Use the **Bin Categorical Field ...** button (or the Expression Builder "bin()" function) to see if you can reduce the number of categories for modeling.
6. Choose a **PValue and Training Subset** to control how model building is done:
 - **PValue**: The PValue is a threshold used in determining if a relationship between an explanatory field and the target is not random and thus should be included in the model. The smaller the PValue, the less likely the relationship is to be based on random variation in the data. We recommend using the default PValue, and then adjusting it as part of the iterative process. As the PValue is increased the model will take longer to run and more factors will be considered – which may create a more robust model as long as the correlation coefficient (% concordance or R^2) remains satisfactory and as long as the now included factors actually make sense.
 - **Training Subset**: You may train your model with a subset of the data. Use the slider to choose the percentage of the model table to randomly use. The total visible rows, rows in the training subset, and what percentage that represents is shown as well. A subset is automatically chosen if the volume of data will make the building time slow. The larger the training subset, the more accurate the model but the longer time it will take to build. You may want to use a smaller subset if you are quickly iterating models, and then use a larger subset for a final training session.
7. **Train Model**: "Training" is the process of creating the model from relationships found in the current data between the explanatory fields and the target field. This can be time consuming! The progress of model building is shown by "Accesses" in the Analytics pane.
8. **Save**: Save the model configuration without training the model.
9. **Cancel**: Close the dialog with no changes.
10. **Help**: Display assistance on using this dialog.

When processing is completed, the model is displayed in two pages in the Analyst.

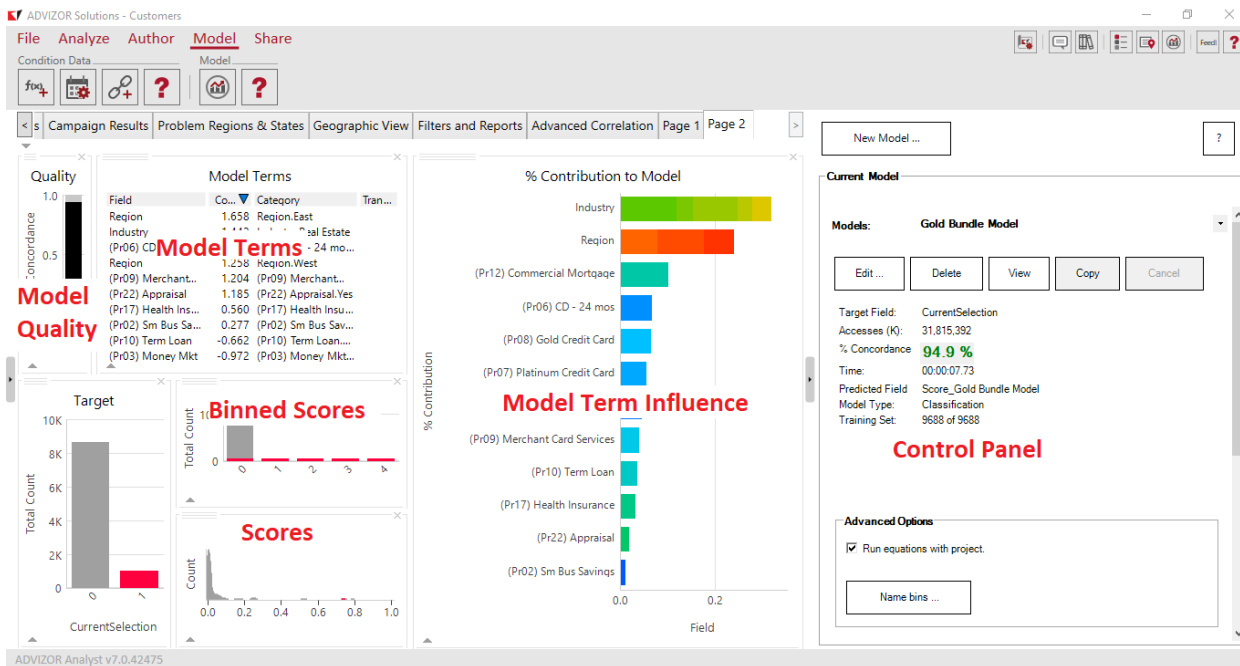
Understanding the Model

Two pages are added to your project with chart that help you understand the relationships in the model. The first page for ordinary regression models with continuous targets shows the information metric, the equation terms, and the relationship of the original target field to the predicted field. The % Contribution to Model bar chart shows what percent of the model's overall predictive ability (given by the "Quality" R² metric) come from each explanatory field.

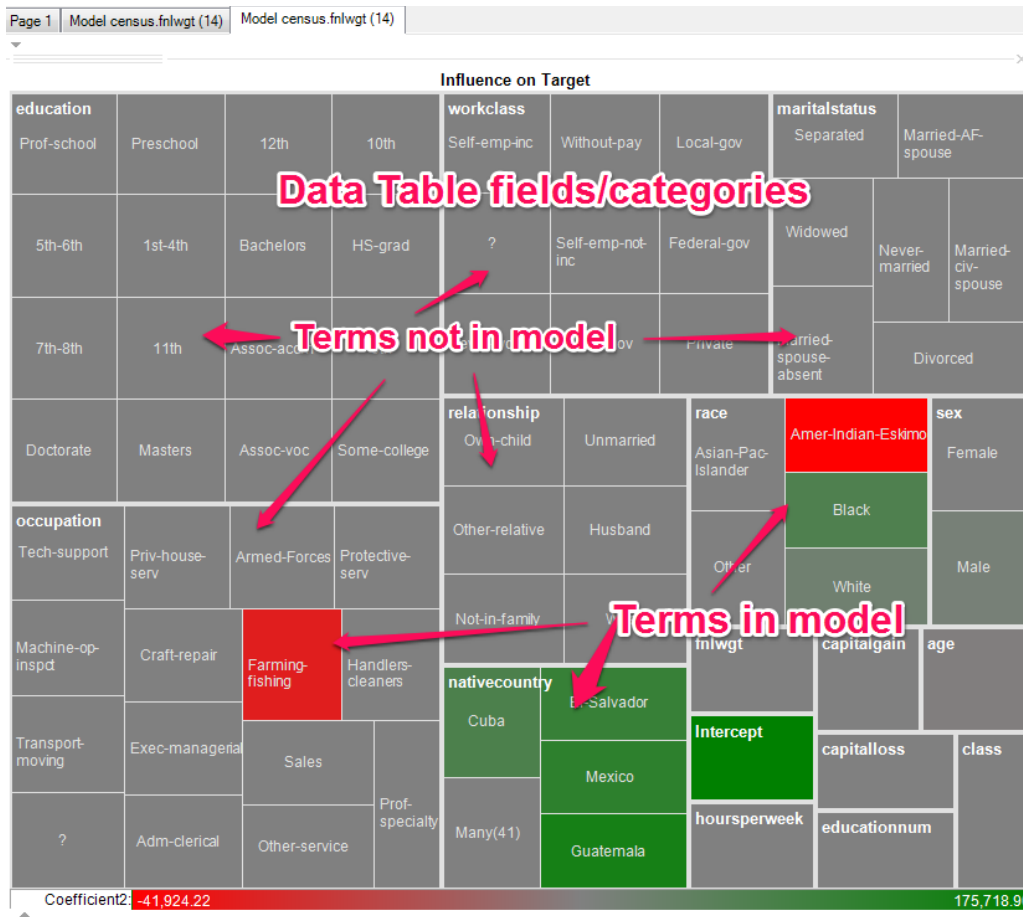
The quality metric accompanies the model and indicates how robust the model is. Linear regression models use the R² value which measures the proportion of variance of the target field explained by the model. Typical values for R² between .7 and .9, though when trying to predict human behavior, even .5 is acceptable. Logistic regression models (aka classification models) use % Concordance which is proportional to the number of predicted values that match the actual value (which for logistic regression is either 1 or 0). The % Concordance typically runs between .6 and .95. For either score if the model quality is too high, it is indicative of overfitting. You may be attempting to use too many factors or using a factor that is identical or a proxy for the target. A low model quality indicates the chosen factors do not predict the target value accurately. Additional factors are needed that can better explain the target.



The first page for classification models shows the information metric, the equation terms, the weight of terms, the original target distribution, the model predicted distribution, the probability scores, and the bins of probability scores. "% Contribution to Mode" bar chart represents how much effect each field that is part of the model has on the target.



A second page contains a Heat Map showing the model tables and which fields influence the target.



Demo Walkthrough

This section will walk through the steps to creating a predictive model within ADVIZOR. It uses the Customer Analysis Demo within ADVIZOR Analyst/X. For ADVIZOR version 7.0 +, you can click on the Customers Demo from the left menu on the start page. For ADVIZOR versions before 7.0, access this project by clicking on the Demo button in the “Getting Started” window that opens when you start ADVIZOR Analyst/X. This project contains data on 10,000 customers of a commercial bank. Information on each customer includes data about the company, such as industry and state, as well as data about the current services the bank provides for each company.

ADVIZOR Solutions

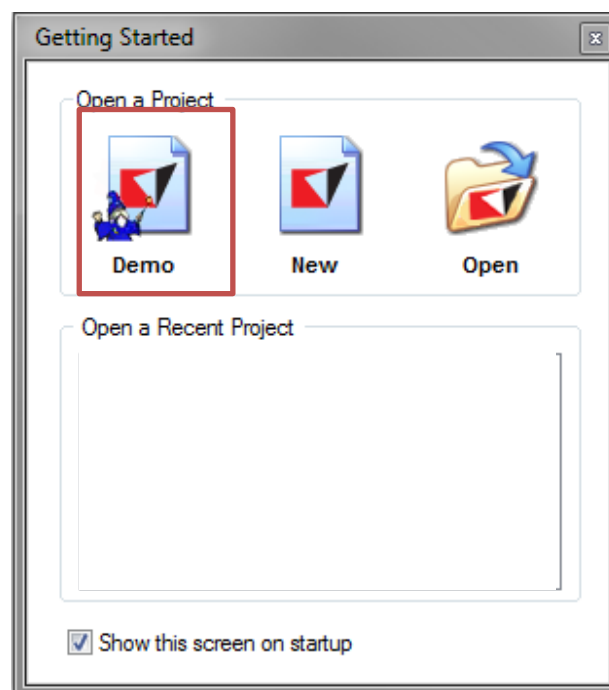
Start

Start a new project...
Open an existing project...

Recent Projects

Demo Projects

Customers
MutualFunds

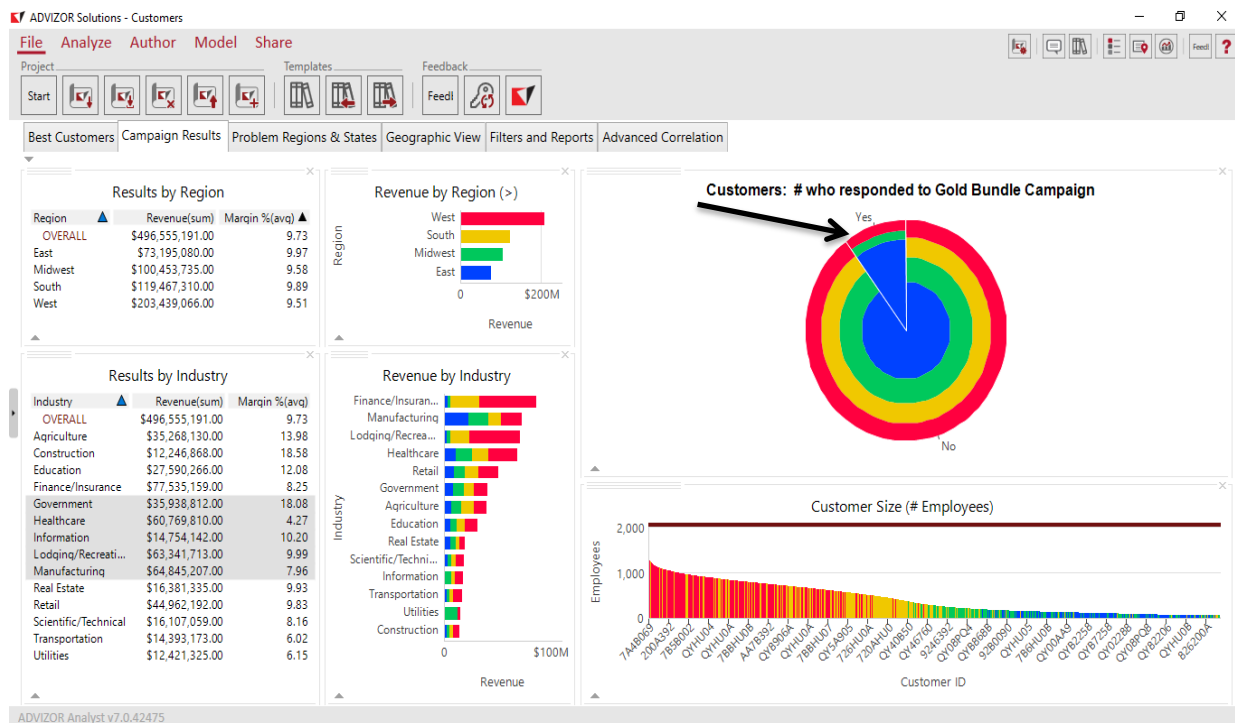


Step 1. Define the business question in terms of data

In the Customer Analysis Demo, there is a field `Response Gold Bundle Campaign`, which shows the response to an advertising campaign. We would like to better understand which other customers would respond to a similar campaign. We want to understand the characteristics of the companies that responded and what, if any, common features are shared among them that distinguishes them from the rest of the bank's customers. The end action item is following up with similar customers with a similar advertising campaign.

Step 2. Select the target field and identify the subset of the data that is relevant to the question (the base population)

This will be a classification model because we will be comparing two groups: those that responded “Yes” to the Gold Bundle Campaign and those that responded “No”. To select the target population, go to the “Campaign Results” page and click on the “Yes” portion of pie chart titled “Customers: # who responded to the Gold Bundle Campaign”. This selects the 984 companies that responded to the campaign.

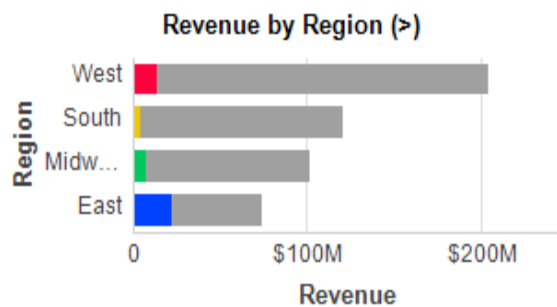
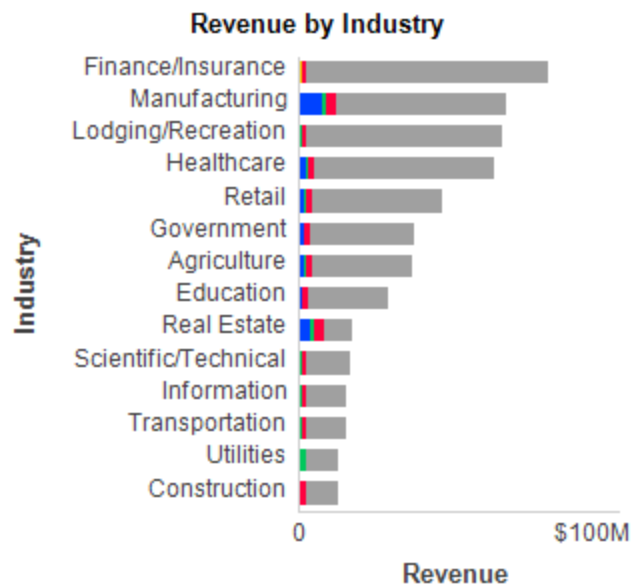
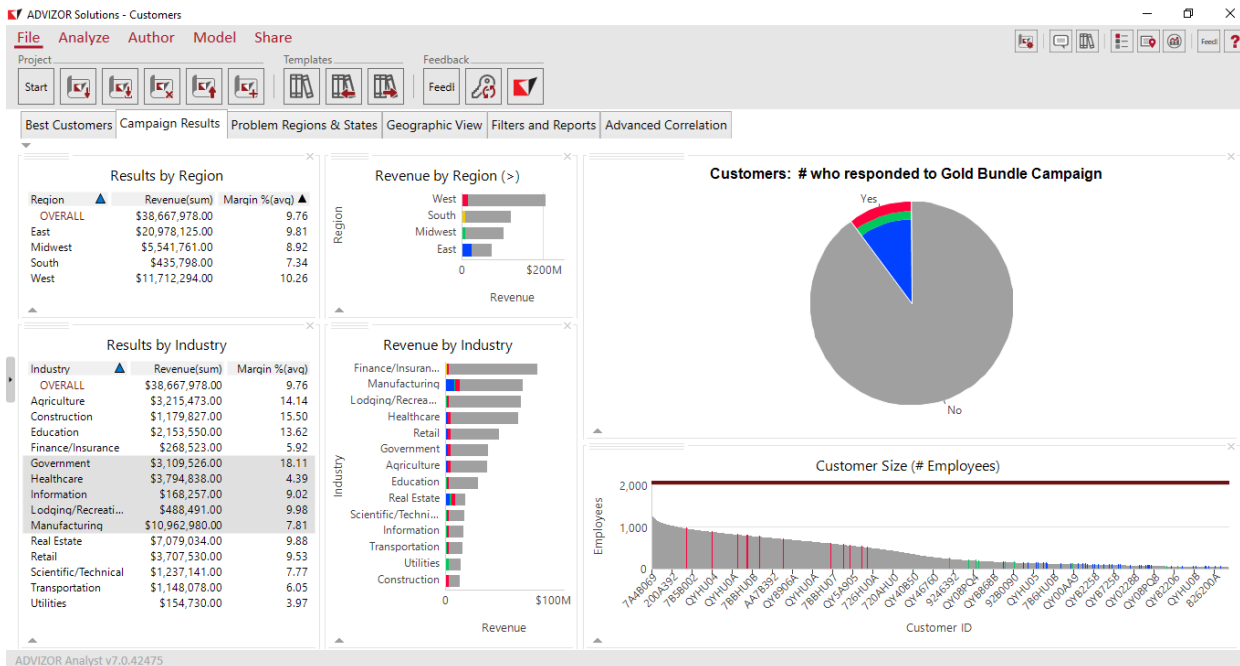


The base population will be the rest of the customers. We could do separate models for each region or for each industry if we expect companies to behave differently in those segments. Additionally, all companies have a response for the Gold Bundle Campaign. Thus, we are not excluding any companies based on the criteria that they didn't get a chance to respond.

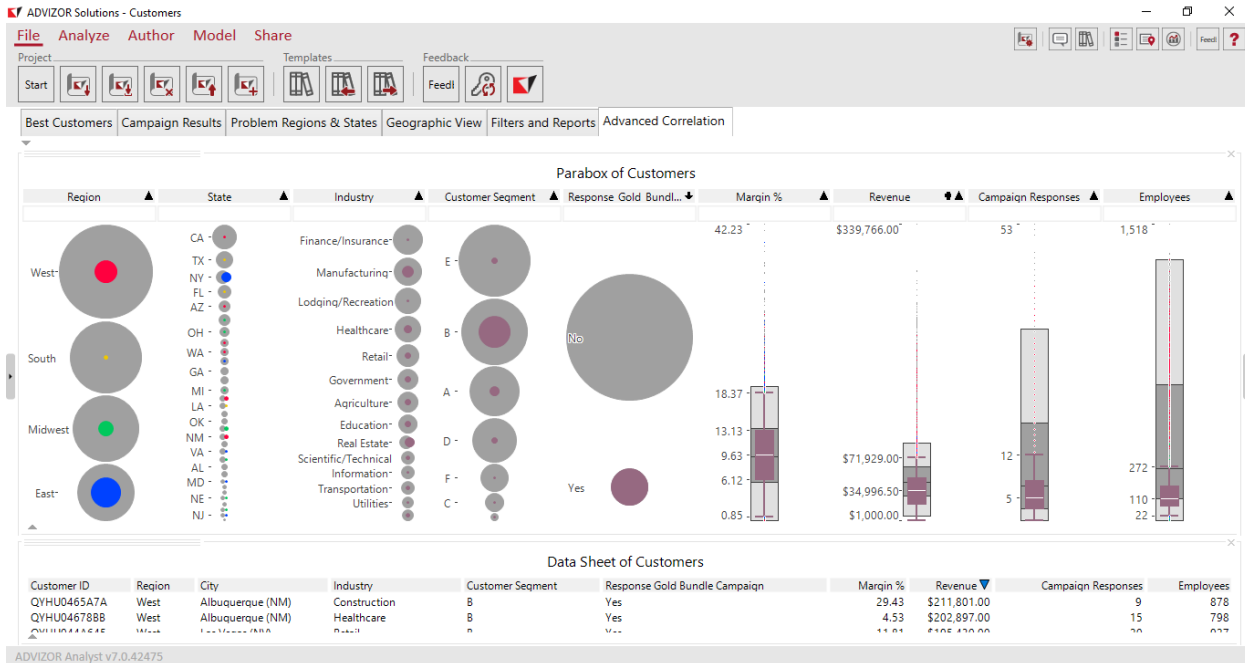
Step 3. Visually Explore and Form Hypotheses

We then look at the selection throughout the rest of the project to see any patterns that stand out from our selection. This can help us decide what possible explanatory factors will go into the model.

Going to the first page, we can see that in Revenue by Region, both West and East have larger proportions selected, suggesting that Region might be an influencing factor. A similar conclusion can be drawn by looking at Industry. It appears that the Gold Bundle campaign performed better with those in the Manufacturing and Real Estate Industries, but poorly with those companies in the Finance/Insurance and Lodging/Recreation industries.



Finally, go to the “Advanced Correlation” page. Displayed here is a Parabox, which shows the distribution of our selection across multiple dimensions. Each column shows a different characteristic of the customer. The colored portion represents our target population. We can see how the target population skews in certain categories. For example, the box plots on the right for Revenue and Employees. We can see the colored portion is lower than the rest, suggesting that smaller companies responded better to the campaign. On the other hand, the distribution for Margin of the selected target companies mirrors that of the grey distribution of Margin for the base population, suggesting that Margin is not a distinguishing characteristic of the target population.

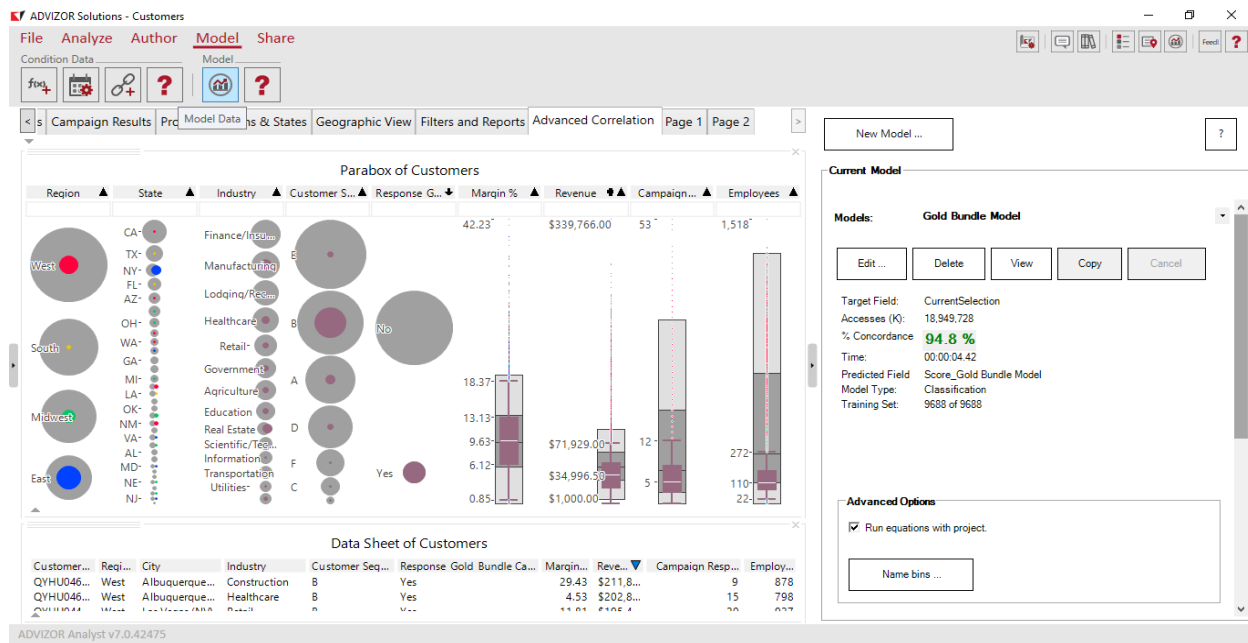


Step 4. Select Explanatory Factors

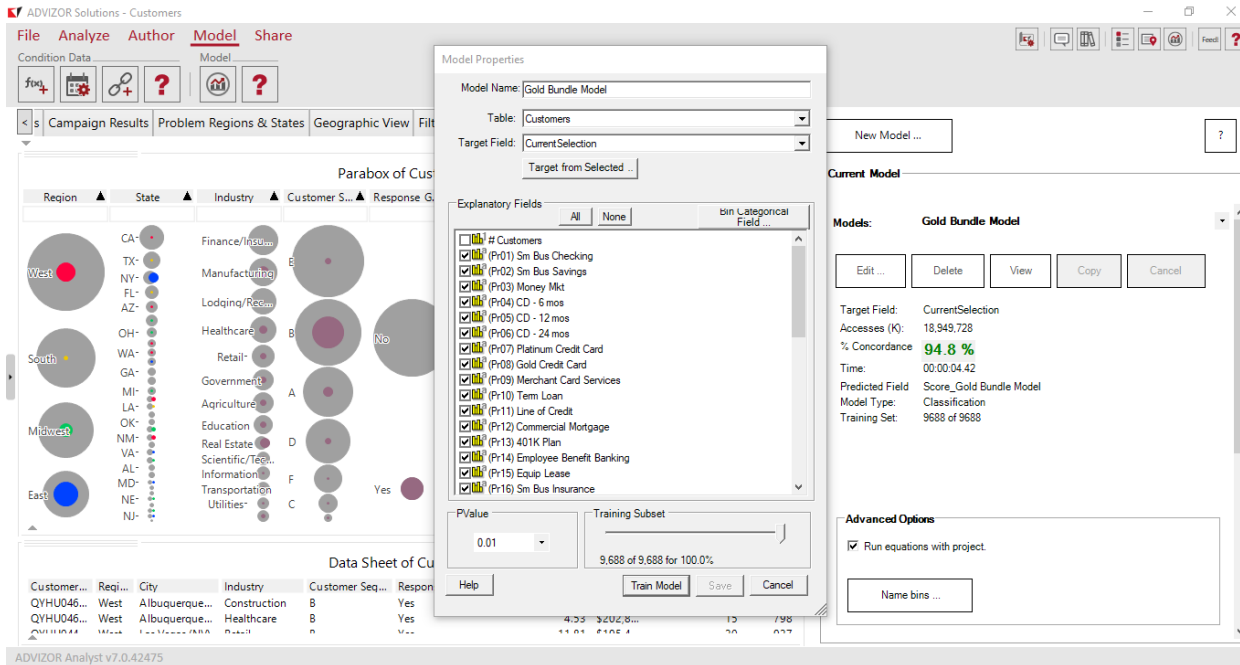
Select our list of explanatory fields by going through the data and identifying fields that we suspect differentiate between our target and base populations. In this project we have some fields, denoted with (Pr), that indicate what products and services the customers have and some fields that describe the characteristics of the customers, such as their industry and size. We might reason that the products and services the customers currently own might influence their response to the “Gold Bundle Campaign”; as such we can include those in the model. This suspected influence does not necessarily have to be a positive correlation; there might be certain products that responders to the “Gold Bundle Campaign” tend to avoid. From our visual exploration, we also found some other factors like Industry and State.

Step 5. Train the model

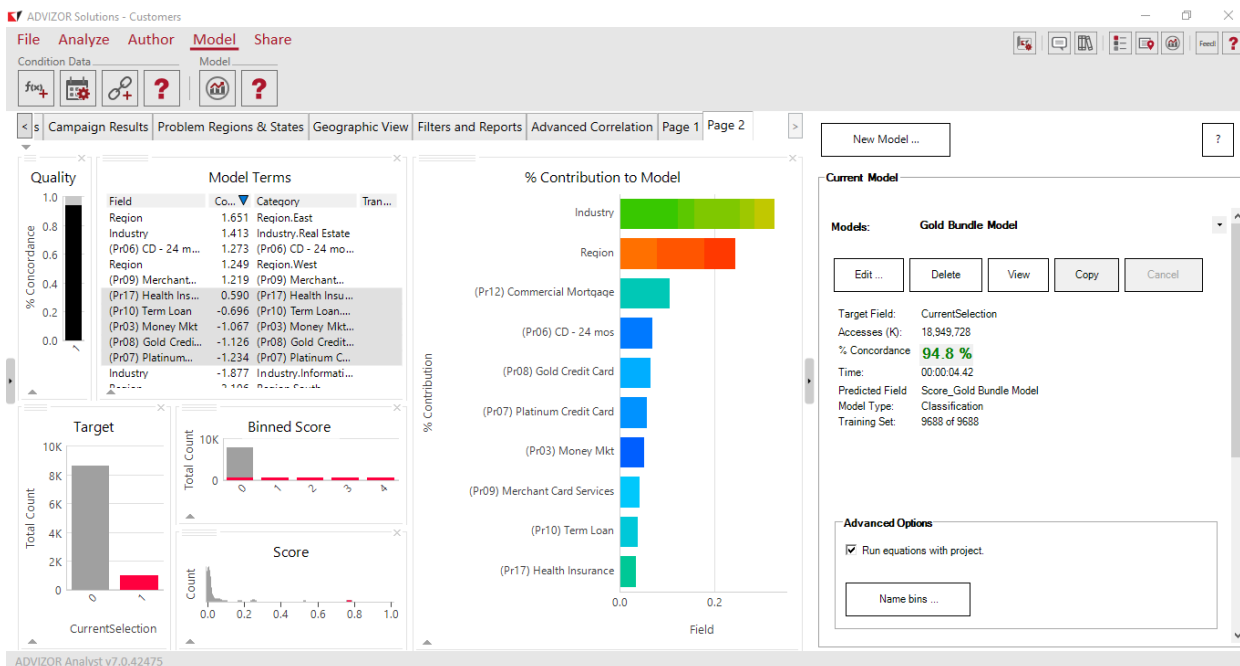
To train the model, open the predictive modeling panel. We can do this by opening the View Menu and clicking on Predictive Model option.



- Click on New Model to create a new model. In the Model Name Field at the top, type in “Gold Bundle Model”
- Keep the Table: as Customers.
- For the Target Field, click on “Target from Selected”. This will make the model run on the current selection, which from Step 2 should be all customers who responded “Yes” to the Gold Bundle Campaign.
- Below, we can choose which fields to include in our model. For now, ensure that the fields that begin with (Pr), Employees, Industry, Margin, Region, and Revenue are the ones checked and the others unchecked. We can leave the other options in the bottom of the window as default.
- When done, click on “Train Model”



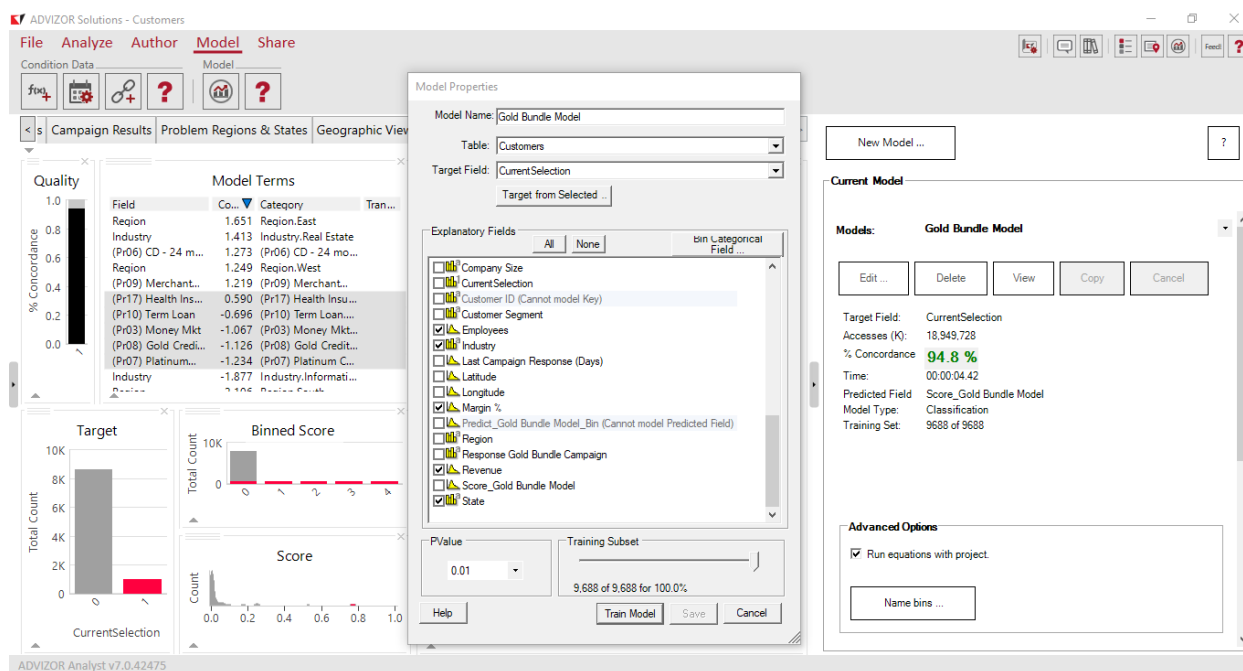
Step 6. Examine the model and iterate



Once the model is built, ADVIZOR will display the results on two pages. On one of the pages, ADVIZOR will show the significant factors, their coefficients, as well as their contribution to the model. We can then look through the list and evaluate what came up as significant with our existing expectations. Are there certain fields that come up as expected? Industry came up as significant and Margin did not come up as significant, both of which we expected. Are there certain fields that come up as surprising? Perhaps one the products we thought would appeal to the responders of the “Gold Bundle Campaign”

turned out to not be correlated. Are there refinements that we can make to the model? For the gold bundle demo, we see that “Region” was one of the factors that came up as significant. Since there are only four “Regions” for all customers, it might be beneficial to look more carefully and at a lower level – the state level. For example, looking at the State column, we see that it is New York that causes the large proportion of “East” and Nevada and New Mexico that causes are larger proportion of the “West” category to be selected while California is significantly underrepresented. Thus, rather than modeling on Region, it would be better to model on State to get the fuller, more nuanced, picture. We can edit and rerun the model to make adjustments to the model.

In the predictive modeling pane, click on the “Edit” button to reopen the model configuration view. Uncheck “Region” and check “State” then Train Model. This will then rebuild the model with the new explanatory factors.

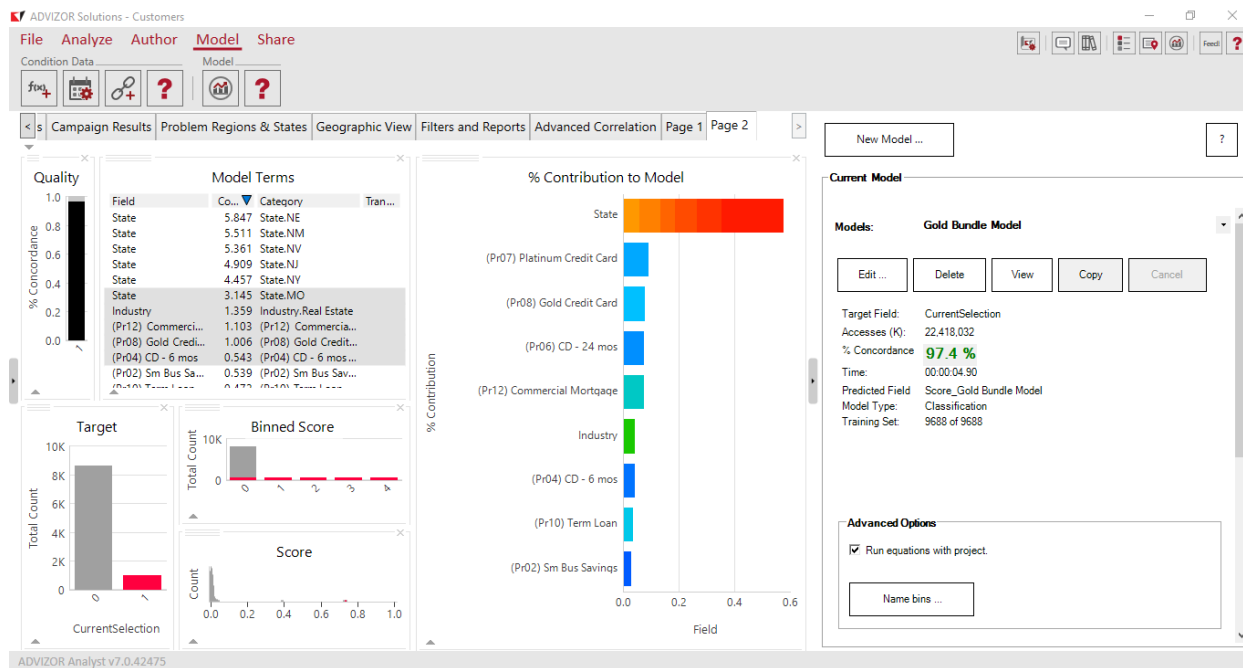


The screenshot displays the ADVIZOR Analyst v7.0.42475 software interface. The main window is titled "ADVIZOR Solutions - Customers" and features a menu bar with "File", "Analyze", "Author", "Model", and "Share". Below the menu bar is a toolbar with icons for various functions. The interface is divided into several panes:

- Model Properties:** A dialog box in the center showing the "Model Name" as "Gold Bundle Model", the "Table" as "Customers", and the "Target Field" as "CurrentSelection". It also includes a "Target from Selected ..." field.
- Model Terms:** A table listing various fields and their corresponding categories. The table has columns for "Field", "Co...", and "Category".
- Explanatory Fields:** A list of fields with checkboxes for selection. The fields include "Company Size", "CurrentSelection", "Customer ID (Cannot model Key)", "Customer Segment", "Employees", "Last Campaign Response (Days)", "Latitude", "Longitude", "Margin %", "Predict_Gold Bundle Model_Bin (Cannot model Predicted Field)", "Region", "Response Gold Bundle Campaign", "Revenue", "Score_Gold Bundle Model", and "State".
- Quality:** A bar chart showing the percentage of concordance for different fields. The y-axis is labeled "% Concordance" and ranges from 0.0 to 1.0.
- Target:** A bar chart showing the total count for different target values. The y-axis is labeled "Total Count" and ranges from 0 to 10K.
- Binned Score:** A bar chart showing the total count for different binned scores. The y-axis is labeled "Total Count" and ranges from 0 to 10K.
- Score:** A line chart showing the score for different values. The x-axis is labeled "Score" and ranges from 0.0 to 1.0.
- Current Model:** A panel on the right showing the current model configuration. It includes buttons for "Edit...", "Delete", "View", "Copy", and "Cancel". It also displays the "Target Field" as "CurrentSelection", the "Accesses (K)" as 18,949,728, the "% Concordance" as 94.8%, the "Time" as 00:00:04.42, the "Predicted Field" as "Score_Gold Bundle Model", the "Model Type" as "Classification", and the "Training Set" as 9688 of 9688.
- Advanced Options:** A section at the bottom right with a checkbox for "Run equations with project" and a button for "Name bins ...".

The bottom status bar indicates the version "ADVIZOR Analyst v7.0.42475".

Step 7. Integrate output into the project



Once the model rebuilds, we see that “State” contributed the most, followed by several products, “Industry”, and several other products. In the Model Terms chart, we can see which states and which industries have an impact on the target behavior. One conclusion that we might draw is that our “Gold Bundle Campaign” resonated with companies in one of the states listed. We can also draw the conclusion that customers in the Real Estate industry also tend to respond positively to the Gold Bundle Campaign. Using the model output pages and seeing which explanatory fields came up a significant helps us to understand the relationships between the target population and the rest of the data. However, in addition to qualitative relationships, once we have a model we are satisfied with, we can use the scores to sort and create lists. Each customer in the table is given a score based on the model results in the range between 0 and 1. This score is in a field denoted “Score_(ModelName)” and can be used in any chart.

For example, lets create a Data Sheet on a new page in the Customers Demo. For the Data Sheet, include the field “Customer ID” and the model score field “Score_(ModelName)”. Clicking on the Model Score Field Header will sort the list of customers in descending order of model score. Thus, we have a sortable list of customers who are most similar to the target field; those that would be most likely to respond to the “Gold Bundle Campaign”.

Step 8. Continually evaluate model performance

After the model is created and a scored list is found, it is important to evaluate the model’s performance in the coming months. Compare the predicted results with the actual outcomes to see how accurate the model is. You can adjust the equation manually if needed or rerun the model with the new data.