# Predictive Analytics

**Course Project**

**Instructions:**

In this project, you will work through a predictive analysis for a decision of your choosing. Ideally your focus will be in an area of your expertise, and you will have access to data on which to base your analysis.

*You will make two submissions in the course of completing this project. After completing part one of the project, you will submit a partially completed version of this project document for instructor review. A submit button can be found on the Part One assignment page.*

*Once you have completed parts two through four of the project, resubmit this project document and any supporting documents to your instructor for grading. A submit button can be found on the Part Four assignment page. Information about the grading rubric is available on any of the course project assignment pages online.* *Do not hesitate to contact your instructor if you have any questions about the project.*

#### **Part One—Identify a Focal Point and a Dependent Variable**

In this part of the course project, you will identify a variable of interest that is critical to a past or pending decision. You will list and describe independent variables that are likely to be associated with the dependent variable. For each independent variable, you will assert its expected impact on the dependent variable based on your familiarity with the context.

In order to satisfy this part of the project, you will identify a target of analysis and a dependent variable, and you will complete the independent variables table.

1. Briefly summarize the decision or prediction that you wish to target using predictive analysis. *(50 words or fewer)*

Optional: If you feel additional description is necessary to capturing the situation, provide a summary of relevant circumstances around the prediction or decision*.* This could include timeline, stakeholders, related outcomes, etc. *(300 words or fewer)*

Target of analysis:

Description of the context:

1. Identify the dependent variable that will guide your prediction or decision.

Dependent variable:

1. Identify at least three independent variables that you believe have association with the dependent variable. For each independent variable, identify it as quantitative or categorical and discuss its expected impact on the dependent variable.

|  |  |  |
| --- | --- | --- |
| **Independent Variable** | | |
| **Summary of independent variable** | **Categorical or quantitative?** | **Argument for / description of the associates with the dependent variable** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**Part Two—Map Decisions to Outcomes**

In this part of the project, you will map how your decision or predicted outcome is related to the independent variables you are considering. You will support this relationship with visualizations.

In order to satisfy this part of the project, you will complete the following.

1. **For each** of the independent variables listed in part 1:
   1. Paste a screenshot of a scatterplot here (include a best fit line for all quantitative variables). Make sure the independent variable is on the horizontal axis and the dependent variable is on the vertical axis.
   2. Write the regression equation of the best fit line in the table below.

|  |  |  |
| --- | --- | --- |
| **Candidate Independent variables** | | |
| **Independent variable** | **Regression equation** | **Screenshot of scatterplot** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

1. **For each** of the independent variables listed in part 1, use plain language to explain why the relationship shown in the scatterplot and regression equation make sense in the context of the situation you’re exploring. However, if either the slope or intercept of the equation seems counterintuitive given your intuition, also make a note of that in your explanation. *(50-100 words per variable)*
2. Run a multiple regression for your data. Paste a screenshot of the results below.

**Part ThreeGenerate a Revised Regression Equation**

In this part of the project, you will refine your regression model potentially by transforming variables for nonlinear relationships and/or including or excluding variables from the regression to address multicollinearity.

In order to satisfy this part of the project, you will complete the following.

Addressing nonlinear relationships:

1. **For each** of the independent variables listed in part 1, paste a screenshot of its residual plot against the dependent variable Y here in your project document.
2. Review the residual plots and determine which, if any, suggest a nonlinear relationship with the dependent variable. List suspected nonlinear variables in the table below.

1. Use the Semilog and Log-log Transforms tool to create a transform scatterplot for each independent variable listed, and include screenshots of the plots in the table.
2. List the independent variables with nonlinear relationships.

|  |  |  |
| --- | --- | --- |
| **Possible nonlinearities** | | |
| **Independent variable** | **Transform used (log or semilog)** | **Screenshot of transform plot** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Addressing Multicollinearity

1. Create a correlation table for your independent variables. Paste a screenshot of your correlation table here.
2. Determine if there are independent variables that may be sources of multicollinearity. List them here with an explanation of why you think they might be a source of multicollinearity.

**Part Four—Validate Your Model**

In this part of the course project, you will make a convincing and minimally technical argument for the validity of your model. In addition, you may test your model using a holdout sample.

In order to satisfy this part of the project, you will complete the following.

1. Share a one-page summary of your project*. (500 words or fewer)*

In your summary, include a brief description of the context and the dependent variable of interest. Make an argument for the viability of your model. Aspects of your argument may be based on statistical details such as p-values for coefficients, signs and magnitudes of coefficients, and R-squared values.

For attributes you have included in your model, be sure to address the consistency of linear relationships and any nonlinearities. Where possible, draw a connection between these attributes and the working realities of the situation being described.

If there are any missing drivers or attributes that have been excluded and which seem relevant to the situation, make a note of these and explain the reasoning and potential impacts of excluding them. In this discussion, be sure to address not only attributes excluded on purpose as well as those excluded because of unavailability of data.