|  |  |  |  |
| --- | --- | --- | --- |
|  | **Python Project Notes** |  | |
|  | A | |  | |
|  | Here is a checklist for things you need to have in your analysis: | |  | |
|  | 1. Data inspection | |  | |
|  | 1. Summary Statistics | |  | |
|  | 1. Correlations | |  | |
|  | 1. Plots, analyzing the data distribution | |  | |
|  | 1. Missing data | |  | |
|  | 1. **Feature Creations (Making Calculated Columns)** | |  | |
|  | 1. **The treatments of non-numeric features** | |  | |
|  | 1. Splitting the training and test sets for the model | |  | |
|  | 1. Training the model | |  | |
|  | 1. Explain, Predict, or Both (Probably prediction, but not explanation) | |  | |
|  | 1. Evaluation | |  | |
|  | * At the end, we want 1 equation that explains all of the data | |  | |
|  |  | |  | |
|  | 1. Data Inspection | |  | |
|  | * Use SHAPE, INFO and HEAD/TAIL commands to inspect the data | |  | |
|  | * Look for | |  | |
|  | * + Missing data | |  | |
|  | * + Non-Numeric variables | |  | |
|  | * + The most promising variables | |  | |
|  | * + Wayts to combine features and create new variables (calculated columns) | |  | |
|  | * + Check for data imbalance, maybe you can drop the imbalanced data | |  | |
|  | * + Think of the types of business problems you can address in a dataset | |  | |
|  |  | |  | |
|  | 1. Summary Statistics | |  | |
|  | * Look for Abnormal values (min, max) | |  | |
|  | * Find the mean values of key variables | |  | |
|  | * Look for outliers that can bias your analylsis | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1. Correlations |  | |
|  | * Use .CORR to generate a correlation matrix | |  | |
|  | * Look for high positive or negative correlations | |  | |
|  | * Look for unexpected correlations in direction and magnitude???? | |  | |
|  |  | |  | |
|  | 1. Plots | |  | |
|  | * (Histogram) | |  | |
|  | * Distribution of you dependent variable (target, y variable) | |  | |
|  | * Distribution of key independent variables | |  | |
|  |  | |  | |
|  | * (Scatter plot) | |  | |
|  | * Relationships between key independent and dependent variables | |  | |
|  |  | |  | |
|  | 1. Missing Data | |  | |
|  | * Option 1 -Dropping the features (columns), (Advised when missing data is >20%) | |  | |
|  | * Option 2 – Drop the observations (rows) | |  | |
|  | * Option 3 – Replace the missing values with the mean or median | |  | |
|  |  | |  | |
|  | 1. Feature Creation | |  | |
|  | * Create new variable | |  | |
|  | * Combine features(columns) to generate informative variables | |  | |
|  | * Ex: Sales per square feet, revenue per sales employee, etc | |  | |
|  |  | |  | |
|  | 1. Non-Numeric Data | |  | |
|  | * Use Label Encoding – Assign each category to a number | |  | |
|  | * One Hot Encoding - ?~~? | |  | |
|  |  | |  | |
|  | 1. Training and Test data split (30%test, 70%training) | |  | |
|  | * Must be random | |  | |
|  | * Can be stratified | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |

|  |  |  |  |
| --- | --- | --- | --- |
|  | * Main Analysis: |  | |
|  | linear Regression for continuous  logistic for binary data regression - This means Choice model, because its between 2 | |  | |
|  | * What are the independent and dependent variables for all of your models? | |  | |
|  | * Why did you choose to use each model? | |  | |
|  | * Make sure that everyone understands what our columns mean. | |  | |
|  | * Perhaps we should use more than 1 model | |  | |
|  |  | |  | |
|  | * Say on 1st slides whether it is explanatory or predictive analysis | |  | |
|  | * Say on 1st slide why is this problem is important | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |
|  |  | |  | |