**D208 Predictive Modeling TIP SHEET**

Here are some great tips and resources for your consideration in completing D208’s Performance Assessment.

1. **Consider using Random Forest in your Performance Assessment:**

<https://www.r-bloggers.com/predict-customer-churn-logistic-regression-decision-tree-and-random-forest/>

1. **If you decide on using R, consider using RVEST and FactoMineR packages.**

**Chapter 5 gives great breakdowns on the tools. You will probably use RVEST and FactoMineR in C744 for your project work. FAMD (Factor Analysis of Mixed Data) contains PCA and MCA methods built-in. Notice that R stops in the Low category so do not bog it down with too much data.**

1. **Study and be able to identify how to treat Mixed Data.**

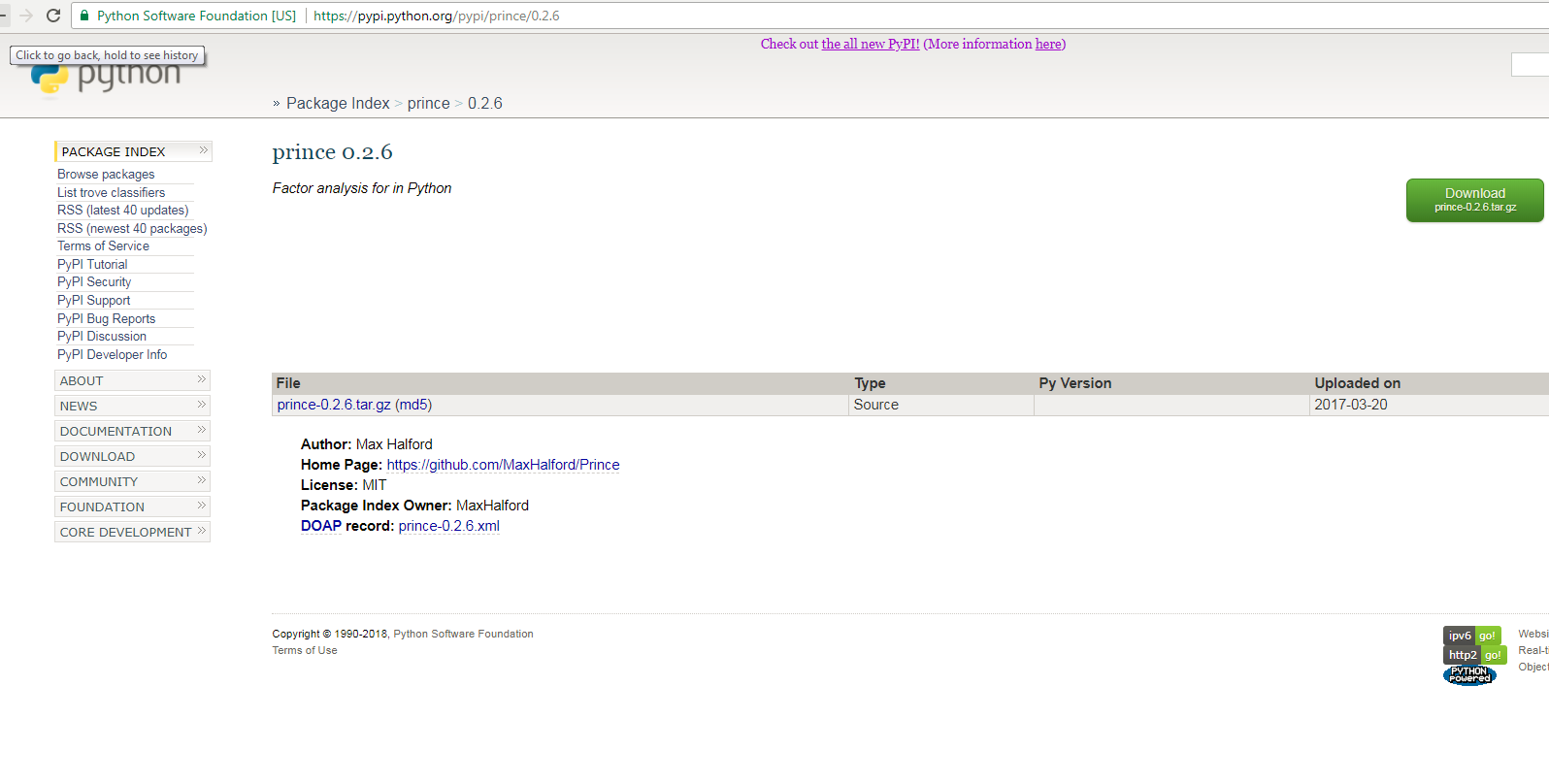
**Recite in your analysis the strengths and weaknesses why you chose the methods you did and state why and how they are better than the alternatives. Discuss the alternative methods.**

1. **Provide at least twenty graphs depicting all phases of the project.**

**Describe and identify each part of your supporting graphics and how they directly relate to the required topic. You must be able to defend your analysis. This is exactly what you will do in the professional arena.**

**Suggestions:**

1. **Univariate Statistics depicting the variable distribution,**
2. **Bivariate Statistics from cleaned and prepared data**
3. **Vector charts depicting Quantitative or Qualitative variables**
4. **Scree plots of dimensionality**
5. **How discriminating data provides relationship among the measures of importance**
6. **Box and whisker plots for Outlier Detection**
7. **If you want to wrangle mixed data in Python instead of R use the Prince package shown below.**



1. **When do you use Logistic Regression in your analysis and Why?**

**Be sure to defend your work by answering this question and say how FAMD plays into your consideration and why. What are its strengths and weaknesses?**

1. **Is Churn the Target Variable in your Study? If so, why? If not, why not?**

**Be able to classify the qualitative and quantitative variables. What is the nominal categorical variable that is your target?**

1. **Identify the Independent Predictor Variables.**

**Be able to classify the predictor variables as nominal or ordinal, qualitative or quantitative.**

1. **What are the goals in manipulating the data?**

**For example: how would you clean the data? Would it need to be split into multiple datasets? After cleaning the data, how would you further treat it in order to make it ready?**

1. **Statistical Identity**

**Can you define the (quantitative) continuous independent variables?**

**What are the categorical (qualitative) independent variables?**

**What is the one categorical dependent variable (target) for the logistic regression?**

**What is the one continuous dependent variable (target) for the multiple linear regression?**

**Is there an identifier for all of the individuals records in the dataset?**

**What is the essential criteria to be predicted?**

1. **Supply screenshots of your code with documentation**

**Use remarks to identify the steps and processing that is going on in your code.**

**Explain your work using flower-boxes and comments in the header and sections.**

**Regards,**

**Dr. William Sewell**

**Course Mentor, IT**