Final Project Instructions

Assignment:

In this situation you are acting as a data science consultant to a bank following a *reduced* CRISP-DM project methodology. The objectives of the assignment are: 1) to predict customer churn for a bank using the machine learning tools discussed in this course and 2) to make an actionable recommendation to the bank articulating what they should consider doing as a business to reduce customer churn. The dataset provides 14 fields as shown below. The training-testing dataset contains 9,000 records in total. The production dataset contains 1,000 records in total.

1. Row Number
2. Customer Id
3. Surname
4. CreditScore
5. Geography
6. Gender
7. Age
8. Tenure
9. Balance
10. Num Of Products

How many accounts, bank account affiliated products the person has

1. Has Cr Card
2. Is Active Member

Subjective, but for the concept

1. Estimated Salary
2. Exited

Did they leave the bank after all? This is the Label

Your assignment is to follow the *reduced* CRISP-DM project methodology provided resulting the selection, building and testing of the appropriate model using Rapidminer using the performance validation method(s) of your choice. The labelled training-testing dataset contains 9,000 records; the unlabeled production dataset contains 1000 records. You should then apply a trained model to the production dataset in order to generate a forecasted label and provide a recommendation follow-up action plan to the bank. You should also examine the variable importance identified in the analysis and provide a recommended course of action for the bank to reduce churn.

By following the *reduced* CRISP-DM methodology provided, you should be sure to check for missing observations, outliers and redundant features. You should consider any relevant omitted variables, and you should explore opportunities to perform feature engineering using existing variables provided in the dataset. You show your work and provide evidence that you have completed the necessary steps sufficiently.

Finally, you should present your findings within a powerpoint document. This should include your business recommendations that are supported by the data. As an appendix to your powerpoint you should supply the scored forecast of the 1000 production observations in rank order of probability of churn.

Note: when making your choice of which machine learning engine to use in production, your priorities should be as follows:

1. Highest Priority=Precision
2. Secondary Priority=Recall
3. Tertiary Priority=Accuracy

You should not consider a modeling engine with an AUC less than 0.60 as a candidate for deployment to production.

As part of your recommendation you should construct a “profile” of customers who are most likely to churn. You should also identify the major market segments that the bank should consider in its marketing to customers.