Southern New Hampshire University

DAT 690 Project Overviews for Final Project

Prompt

From the pilot work we did with the business on credit, attrition, and churn, the business users have requested a deliverable of one project for each business area. These projects will need to be modeled and implemented into production.

Project Instructions

You will need to perform variable analysis to identify variables that are important. You can do this by using descriptive statistics and running clustering and tree models, as well as PCA. If needed, you can also derive variables from the data provided.

Once you have an initial set of variables, continue with your model design using both identified models above to compare and contrast how the models perform. Select your final variables and final model design, and then baseline the results. As part of your turnover to production, you will be provided a new verification file, which you will then run through your final model to score against your baseline. The final step will be to complete your turnover form for IT to deploy your model to production.

Business Focuses

Your team will need to **select one of the projects below to complete.**

Business Focus	Projects
Credit	Loan Default
	Credit Amount
Attrition	Employee Attrition
	Employee Salary
Churn	Customer Churn
	Customer Revenue

Credit

Credit Project One: Loan Default

The first option for the credit project is the prediction of loan default. This project uses the same data set as the pilot and is answering the same question to predict the likelihood of default. This project is now a fully developed model that will be deployed to production as part of the normal business process. The production model will need to have proper factor analysis and variable reduction. Recall that less is typically better for predictions. The final production model to be used is either Naïve Bayes or Logistic Regression. The accuracy of predicting the likelihood of default should be in the 70 percent range. You are provided a sample file to use to train and test your model. Ensure that you properly set the data types of your variables (for example, categorical and numerical).

Credit Project Two: Credit Amount

The second option for the credit project is the prediction of the amount of credit to extend. This is a new data set that contains over 100 variables and 5,000 rows of data. For this model, you will use a predictive model of either General Linear Regression or Basic Neural Net to predict the amount of credit to extend. You are provided a sample file to use for your model. Ensure that you properly set the data types of your variables (for example, categorical and numerical).

Southern New Hampshire University

Attrition

Attrition Project One: Employee Attrition

The first option for the attrition project is the prediction of employee attrition. This project uses the same data set as the pilot and is answering the same question to predict the likelihood of attrition. This project is now a fully developed model which will be deployed to production as part of the normal business process. The production model will need to have proper factor analysis and variable reduction. Recall that less is typically better for predictions. The final production model to be used is either Naïve Bayes or Logistic Regression. The accuracy of predicting the likelihood of default should be in the 70 percent range. You are provided a sample file to use to train and test your model. Ensure that you properly set the data types of your variables (for example, categorical and numerical).

Attrition Project Two: Employee Salary

The second option for the attrition project is the prediction of the salary an employee would desire to keep them from leaving. HR will be able to review expected salary to current salary as part of their employee attrition reviews. This is a new salary data set that is joined with the original pilot data. For this model, you will use a predictive model of either General Linear Regression or Basic Neural Net to predict the salary of an employee. You are provided a sample file to use for your model. Ensure that you properly set the data types of your variables (for example, categorical and numerical).

Churn

Churn Project One: Customer Churn

The first option for the churn project is the prediction of customer churn. This project uses the same data set as the pilot and is answering the same question to predict the likelihood of churn. This project is now a fully developed model which will be deployed to production as part of the normal business process. The production model will need to have proper factor analysis and variable reduction. Recall that less is typically better for predictions. The final production model to be used is either Naïve Bayes or Logistic Regression. The accuracy of predicting the likelihood of default should be in the 70 percent range. You are provided a sample file to use to train and test your model. Ensure that you properly set the data types of your variables (for example, categorical and numerical).

Churn Project Two: Customer Revenue

The second option for the churn project is the prediction of the revenue a customer generates based on usage inputs. This is a subset of the original pilot data. It will be used by the marketing team to determine if it is worth retaining a churn customer. For this model, you will use a predictive model of either General Linear Regression or Basic Neural Net to predict the amount of revenue the customer generates. You are provided a sample file to use for your model. Ensure that you properly set the data types of your variables (for example, categorical and numerical).