NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

# PREDICT 454: Team Project Guide

## Getting Started

Everyone in Predict 454 will have to participate in a team project. Each team will have four to six members, and each team will undertake a 'large' modeling effort in the form of a team project. The objectives of the team project are to provide you with an unstructured problem in a learning environment that is similar to the better work environments in industry, after all you might be lucky to be on a team of the quality of your classmates. This document will outline the project regulations and guidelines, and the approved project topics.

## Approved Projects

The second half of this document will outline several approved team project topics. Each of these topics is of the form of a competition data set such as the data provided in a typical Kaggle competition. Other projects could be approved on an individual basis provided that there is enough interest to form a team (four to six people), provided that the data is deemed to be of sufficient size and quality, provided that the data is not extensively used as an internet tutorial, and provided that you are able to define appropriate predictive problems. In the past students have been interested in working on investment projects and sports analytics projects. These projects can be approved if you can meet the demonstration conditions.

For example Kaggle has a widely popular tutorial on modeling the survivors of the Titanic. While I encourage all students in the MSPA program to work through this tutorial, this data set is not appropriate for Predict 454 since it is an online tutorial.

## Obtaining Project Data Sets

The data sets and the associated auxiliary files for the projects presented in this document can be downloaded from my NU Depot file sharing account.

https://northwestern.box.com/s/usqhktc4acwagv7j47pt

## General Paper Guidelines

The general guidelines are as follows.

(1) The paper should follow the appropriate final paper format (one of the two possible formats outlined in this project guide).

(2) The paper should be well written. The formats are designed to help you with the overall paper structure. The remainder of the presentation is left to you.

(3) The term data is the plural of datum. We should write 'the data are', not 'the data is'.

(4) The body of the paper should be between twenty and forty pages, double spaced. The title page, the bibliography, and any relevant appendices do not count against the page count. A good target number is the midpoint of thirty pages.

(5) Graphics and tables should be labeled so that they can be referenced. Typical labels are 'Figure 1' and 'Table 1'. Graphics and tables should also be large enough to be easily read and centered in the page.

## General Grading Guidelines

Your final paper is worth a total of 200 points. Here is how those points will be broken down.

(1) Paper Formulation and Problem Statement (50 points)
- This category includes the 'Introduction', 'The Modeling Problem', and the 'Literature Review' (if relevant). It should be very clear to the reader what your problem is, and perhaps why it is important.
- Your Introduction should frame your paper.
- Your Problem Statement should frame your problem.
- Your Literature Review should focus on previous work that is of importance to your work.

(2) Exploratory Data Analysis (25 points)
- This category will include the description of your data and your exploratory data analysis.
- You should only show relevant EDA results.
- If you do not feel that a result is worth discussing, then it should not be included in your report.

(3) Predictive Modeling (50 points)
- This category will focus on the section 'Predictive Modeling: Methods and Results'.
- Be sure to present each method as its own subsection.
- Each subsection should be clear as to what model you are fitting and what results you obtained.

(4) Comparison of Results (25 points)
- A summary of your results should be presented in a format that makes the effectiveness of the different models easily comparable.

(5) Conclusion (25 points)
- The 'Conclusions' section should wrap up your paper. It should not be two or three sentences long.
- The Conclusion can include discussion of what worked well and what did not, and suggest other avenues of investigation.

(6) Overall Exposition (25 points)
- This final report will be graded for exposition. The report should be well written and use correct English grammar, punctuation, and spelling.
- Each member of the team should proofread that report. In a team project there is no excuse for typos, poor grammar, or poor spelling. .
- Paper should read well and be properly formatted with respect to the specific and general guidelines.

## Final Paper Format – Competition Data

Here is an outline of a suggested format for a team project that uses a competition style data set, i.e. a data set that comes in completion with a stated objective. Each section is listed by name in the order that it should appear in the paper.

**1. Introduction**

- Provide an overview and general statement of the problem.

- Provide a general discussion of how you have approached the problem and highlight some of the interesting results.

- Remember that the Introduction is an introduction to your paper.

**2. The Modeling Problem**

- Provide the initial technical discussion of the problem with a general description of the data.

- This section will inform an intelligent reader, i.e. a reader trained in modeling, of all of the technical details needed to better follow your paper. A reader should not have to read twenty pages deep into your paper to get a basic understanding of your problem, hence we have this section in the paper. A technical reader should be able to read this section and then know what to anticipate in the remainder of the paper.

**3. The Data**

- Provide a description of the data.

- If you have a small number of variables, then a table data dictionary of these variables would be a good choice. If you have a large number of predictor variables, then describe the response variable and the remainder of the data in the most appropriate format.

- Provide a general description the data and any relevant material that is interesting or helpful.

**4. Exploratory Data Analysis**

4a. Traditional EDA

- Provide highlights of important features in your data.

- Discuss how these features may affect (or have affected) your modeling choices.

4b. Model Based EDA

- If relevant, fit a naive model and use it to highlight important features in your data.

- In general you will want to discuss any transformations that the EDA may have prompted you to make.

**5. Predictive Modeling: Methods and Results**

5a. The Train/Test Data

- Provide a brief description of your train/test split or your k-fold cross validation.

5.x Each Individual Model

- You should have a suite of final models for comparison. These final models should use different statistical techniques whenever possible, or simply be fundamentally different if the same technique is used.

- Each model should have its own subsection with the model details. Model details should include the actual model and a GOF analysis on the training data set.

- Model results should be presented in a table. Model results should not be cut-and-pasted R output.

**6. Comparison of Results**

- This project is a project about predictive modeling. We should evaluate each predictive model by a metric for predictive accuracy both in-sample and out-of-sample. This comparison should be easy for the reader to understand. It probably should include both plots and tables.

- If you try to perform any model averaging or committee machines, then the discussion of the averaging technique and the results should be its own subsection in this section. If you use a technique like random forest, it should simply be treated as a standard method (i.e. presented in Section 5).

**7. Conclusions**

- Conclude your paper. Reiterate your problem and highlight your results.

- How would you characterize the overall quality of your results?

- Do you have any recommendations for approaching the problem in a different manner or with different techniques? Would you recommend any particular avenues for future research?

- A lot of times a good conclusion reads like a good abstract.

**8. Bibliography (if needed)**

- References can be constructed using any valid style format. However, references can be cited in your paper using a name-year citing or by using the number scheme, e.g. Bhatti [1].

- One type of citing used in the program is the APA style format. However, it can be difficult to use with some types of sources, hence we will allow the number format for convenience.

- When citing references consider using: Bhatti [1] for a single author, Bhatti and Lucas [2] for two authors, and Bhatti et. al. [3] for three or more authors.

**X. Appendices**

- Appendices are where we put useful, but not primary information. Useful but not primary information could be summary statistics, auxiliary plots, or any additional or auxiliary details and discussion.

## Final Paper Format – Non-Competition Data

Here is an outline of a suggested format for a team project that does not use a competition style data set, i.e. a data set that comes in completion with a stated objective. Typically this format will be of use to teams that are trying to reproduce a research study and teams that have designed their own research study.

It is difficult to provide an exact paper template for a non-competition data project. Make any needed adjustments to optimize your presentation.

Each section is listed by name in the order that it should appear in the paper.

**1. Introduction**

- Provide an overview and general statement of the problem.

- Provide a general discussion of how you have approached the problem and highlight some of the interesting finds.

- Remember that the Introduction is an introduction to your paper.

**2. The Modeling Problem**

- Provide the initial technical discussion of the problem with a general description of the data.

- This section will inform an intelligent reader, i.e. a reader trained in modeling, of all of the technical details needed to better follow your paper. A reader should not have to read twenty pages deep into your paper to get a basic understanding of your problem, hence we have this section in the paper. A technical reader should be able to read this section and then know what to anticipate in the remainder of the paper.

**3. Literature Review**

- Document your information sources. How did previous researchers go about studying this topic, and what were their results?

**4. The Data**

- Provide a description of the data.

- If you have a small number of variables, then a table data dictionary of these variables would be a good choice. If you have a large number of predictor variables, then describe the response variable and the remainder of the data in the most appropriate format.

- Provide a general description the data and any relevant material that is interesting or helpful.

## 5. Exploratory Data Analysis

5a. Traditional EDA

- Provide highlights of important features in your data.

- Discuss how these features may affect (or have affected) your modeling choices.

5b. Model Based EDA

- If relevant, fit a naive model and use it to highlight important features in your data.

- In general you will want to discuss any transformations that the EDA may have prompted you to make.

## 6. Predictive Modeling: Methods and Results

6a. The Train/Test Data

- Provide a brief description of your train/test split or your k-fold cross validation.

6.x Each Individual Model

- You should have a suite of final models for comparison. These final models should use different statistical techniques whenever possible, or simply be fundamentally different if the same technique is used.

- Each model should have its own subsection with the model details. Model details should include the actual model and a GOF analysis on the training data set.

- Model results should be presented in a table. Model results should not be cut-and-pasted R output.

## 7. Comparison of Results

- This project is a project about predictive modeling. We should evaluate each predictive model by a metric for predictive accuracy both in-sample and out-of-sample. This comparison should be easy for the reader to understand. It probably should include both plots and tables.

- If you try to perform any model averaging or committee machines, then the discussion of the averaging technique and the results should be its own subsection in this section. If you use a technique like random forest, it should simply be treated as a standard method (i.e. presented in Section 5).

**8. Conclusions**

- Conclude your paper. Reiterate your problem and highlight your results.

- How would you characterize the overall quality of your results?

- A lot of times a good conclusion reads like a good abstract.

- Do you have any recommendations for approaching the problem in a different manner or with different techniques? Would you recommend any particular avenues for future research?

- A lot of times a good conclusion reads like a good abstract.

**9. Bibliography (if needed)**

- References can be constructed using any valid style format. However, references can be cited in your paper using a name-year citing or by using the number scheme, e.g. Bhatti [1].

- One type of citing used in the program is the APA style format. However, it can be difficult to use with some types of sources, hence we will allow the number format for convenience.

- When citing references consider using: Bhatti [1] for a single author, Bhatti and Lucas [2] for two authors, and Bhatti et. al. [3] for three or more authors.

**X. Appendices**

- Appendices are where we put useful, but not primary information. Useful but not primary information could be summary statistics, auxiliary plots, or any additional or auxiliary details and discussion.

## ADZUNA Job Salary Prediction

This project is a former Kaggle competition. Here is the problem description from the Kaggle website.

https://www.kaggle.com/c/job-salary-prediction

### Problem Description

ADZUNA wants to build a prediction engine for the salary of any UK job ad, so they can make huge improvements in the experience of users searching for jobs, and help employers and jobseekers figure out the market worth of different positions. At the moment, approximately half of the UK job ads they index have a salary publicly displayed. They need your help to bring more transparency to this important market.

Adzuna has a large dataset (hundreds of thousands of records), which is mostly unstructured text, with a few structured data fields. These can be in a number of different formats because of the hundreds of different sources of records. Adzuna needs the help of the Kaggle community to figure out the best techniques to apply to this data set to build a highly accurate predictive model for new ads. You will build, train and test your salary prediction engines against a wide field of competitors.

As an added perk, Adzuna intends to implement their chosen model on their website, both in the UK and worldwide. You will have the satisfaction of seeing your work implemented in production and change the way people search for jobs in the future.

Successful models will incorporate some analysis of the impact of including different keywords or phrases, as well as making use of the structured data fields like location, hours or company. Some of the structured data shown (such as category) is 'inferred' by Adzuna's own processes, based on where an ad came from or its contents, and may not be "correct" but is representative of the real data.

You will be provided with a training data set on which to build your model, which will include all variables including salary. A second data set will be used to provide feedback on the public leaderboard. After approximately 6 weeks, Kaggle will release a final data set that does not include the salary field to participants, who will then be required to submit their salary predictions against each job for evaluation.

## Allstate Purchase Prediction Challenge

This project is a former Kaggle competition. Here is the description from the Kaggle website.

https://www.kaggle.com/c/allstate-purchase-prediction-challenge

### Problem Description

As a customer shops an insurance policy, he/she will receive a number of quotes with different coverage options before purchasing a plan. This is represented in this challenge as a series of rows that include a customer ID, information about the customer, information about the quoted policy, and the cost. Your task is to predict the purchased coverage options using a limited subset of the total interaction history. If the eventual purchase can be predicted sooner in the shopping window, the quoting process is shortened and the issuer is less likely to lose the customer's business.

Using a customer's shopping history, can you predict what policy they will end up choosing?

# COIL 2000 – Caravan Insurance Challenge

An insurance company wants to know who their customers are. This problem is similar to the Allstate prediction problem except it is simplified by not questioning the policy type. This is a very common targeted marking problem used in almost every industry.

http://kdd.ics.uci.edu/databases/tic/tic.html

## Problem Description

For the COIL Competition there were two tasks:

- Predict which customers are potentially interested in a caravan insurance policy.

- Describe the actual or potential customers; and possibly explain why these customers buy a caravan policy.

### *Prediction*

We want you to predict whether a customer is interested in a caravan insurance policy from other data about the customer. Information about customers consists of 86 variables and includes product usage data and socio-demographic data derived from zip area codes. The data was supplied by the Dutch data mining company Sentient Machine Research and is based on a real world business problem. The training set contains over 5000 descriptions of customers, including the information of whether or not they have a caravan insurance policy. A test set contains 4000 customers.

For the prediction task, the underlying problem is to the find the subset of customers with a probability of having a caravan insurance policy above some boundary probability. The known policyholders can then be removed and the rest receives a mailing. The boundary depends on the costs and benefits such as of the costs of mailing and benefit of selling insurance policies. To approximate this problem, we want you to find the set of 800 customers in the test set that contains the most caravan policy owners.

### *Description*

The purpose of the description task is to give a clear insight to why customers have a caravan insurance policy and how these customers are different from other customers. Descriptions can be based on regression equations, decision trees, neural network weights, linguistic descriptions, evolutionary programs, graphical representations or any other form. of solutions (e.g. minimize a loss function, maximize comprehensibility, minimize response time, etc.)?

The descriptions and accompanying interpretation must be comprehensible, useful and actionable for a marketing professional with no prior knowledge of computational learning technology. The value of a description is inherently subjective.

## Customer Relationship Management: Modeling Churn, Appetency, and Up-Sell

This project is the KDD Cup 2009: Customer Relationship Prediction. Here is the problem description as provided by SIGKDD.

http://www.sigkdd.org/kdd-cup-2009-customer-relationship-prediction

### Problem Description

**KDD Cup 2009: Overview - This Year's Challenge**

Customer Relationship Management (CRM) is a key element of modern marketing strategies. The KDD Cup 2009 offers the opportunity to work on large marketing databases from the French Telecom company Orange to predict the propensity of customers to switch provider (churn), buy new products or services (appetency), or buy upgrades or add-ons proposed to them to make the sale more profitable (up-selling).

The most practical way, in a CRM system, to build knowledge on customer is to produce scores. A score (the output of a model) is an evaluation for all instances of a target variable to explain (i.e. churn, appetency or up-selling). Tools which produce scores allow to project, on a given population, quantifiable information. The score is computed using input variables which describe instances. Scores are then used by the information system (IS), for example, to personalize the customer relationship. An industrial customer analysis platform able to build prediction models with a very large number of input variables has been developed by Orange Labs. This platform implements several processing methods for instances and variables selection, prediction and indexation based on an efficient model combined with variable selection regularization and model averaging method. The main characteristic of this platform is its ability to scale on very large datasets with hundreds of thousands of instances and thousands of variables. The rapid and robust detection of the variables that have most contributed to the output prediction can be a key factor in a marketing application.

The challenge is to beat the in-house system developed by Orange Labs. It is an opportunity to prove that you can deal with a very large database, including heterogeneous noisy data (numerical and categorical variables), and unbalanced class distributions. Time efficiency is often a crucial point. Therefore part of the competition will be time-constrained to test the ability of the participants to deliver solutions quickly.

# Forest Cover Type

This is a pure prediction problem in the classic military/defense scenario. The problem is a pure prediction problem, and it has the nice student learning feature of having enough, but not too many, predictor variables. Students interested in using machine learning techniques like neural networks, random forests, and support vector machines will want to consider this topic.

## Problem Description

Predict forest cover type for 30 x 30 meter cells obtained from US Forest Service (USFS) Region 2 Resource Information System (RIS) data. The objective is to predict the forest cover type as a multiclass classification problem based on the associated attributes (features).

http://kdd.ics.uci.edu/databases/covertype/covertype.html

## Data Characteristics

The actual forest cover type for a given observation (30 x 30 meter cell) was determined from US Forest Service (USFS) Region 2 Resource Information System (RIS) data. Independent variables were derived from data originally obtained from US Geological Survey (USGS) and USFS data. Data is in raw form (not scaled) and contains binary (0 or 1) columns of data for qualitative independent variables (wilderness areas and soil types).

### *Summary Statistics*

| | |
|---|---|
| Number of instances (observations) | 581012 |
| Number of Attributes | 54 |
| Attribute breakdown | 12 measures, but 54 columns of data (10 quantitative variables, 4 binary wilderness areas and 40 binary soil type variables) |
| Missing Attribute Values | None |

### *Variable Information*

Given is the variable name, variable type, the measurement unit and a brief description. The forest cover type is the classification problem. The order of this listing corresponds to the order of numerals along the rows of the database.

| Name | Data Type | Measurement | Description |
|---|---|---|---|
| Elevation | quantitative | meters | Elevation in meters |
| Aspect | quantitative | azimuth | Aspect in degrees azimuth |
| Slope | quantitative | degrees | Slope in degrees |
| Horizontal_Distance_To_Hydrology | quantitative | meters | Horz Dist to nearest surface water features |
| Vertical_Distance_To_Hydrology | quantitative | meters | Vert Dist to nearest surface water features |
| Horizontal_Distance_To_Roadways | quantitative | meters | Horz Dist to nearest roadway |
| Hillshade_9am | quantitative | 0 to 255 index | Hillshade index at 9am, summer solstice |
| Hillshade_Noon | quantitative | 0 to 255 index | Hillshade index at noon, summer soltice |
| Hillshade_3pm | quantitative | 0 to 255 index | Hillshade index at 3pm, summer solstice |
| Horizontal_Distance_To_Fire_Points | quantitative | meters | Horz Dist to nearest wildfire ignition points |
| Wilderness_Area (4 binary columns) | qualitative | 0 (absence) or 1 (presence) | Wilderness area designation |
| Soil_Type (40 binary columns) | qualitative | 0 (absence) or 1 (presence) | Soil Type designation |
| Cover_Type (7 types) | integer | 1 to 7 | Forest Cover Type designation |

## *Code Designations*

```
Wilderness Areas:
    1 -- Rawah Wilderness Area
    2 -- Neota Wilderness Area
    3 -- Comanche Peak Wilderness Area
    4 -- Cache la Poudre Wilderness Area

Soil Types:
    1 to 40 : based on the USFS Ecological Landtype Units for this study area.

Forest Cover Types:
    1 -- Spruce/Fir
    2 -- Lodgepole Pine
    3 -- Ponderosa Pine
    4 -- Cottonwood/Willow
    5 -- Aspen
    6 -- Douglas-fir
    7 -- Krummholz
```

## *Class Distribution*

```
Number of records of Spruce-Fir:         211840
Number of records of Lodgepole Pine:     283301
Number of records of Ponderosa Pine:      35754
Number of records of Cottonwood/Willow:    2747
Number of records of Aspen:                9493
Number of records of Douglas-fir:         17367
Number of records of Krummholz:           20510
Number of records of other:                   0

Total records:                           581012
```

## Loan Default Prediction – Imperial College London

This focuses on the default prediction and loss severity problem. In this problem you will have to predict if a person will default (in the scoring classifier sense) and also predict the size (or severity) of the loss given the default. It is a very common loss modeling problem in the banking industry.

https://www.kaggle.com/c/loan-default-prediction/data

Here is the problem description from Kaggle.

### Problem Description

This data corresponds to a set of financial transactions associated with individuals. The data has been standardized, de-trended, and anonymized. You are provided with over two hundred thousand observations and nearly 800 features. Each observation is independent from the previous.

For each observation, it was recorded whether a default was triggered. In case of a default, the loss was measured. This quantity lies between 0 and 100. It has been normalised, considering that the notional of each transaction at inception is 100. For example, a loss of 60 means that only 40 is reimbursed. If the loan did not default, the loss was 0. You are asked to predict the losses for each observation in the test set.

Missing feature values have been kept as is, so that the competing teams can really use the maximum data available, implementing a strategy to fill the gaps if desired. Note that some variables may be categorical (e.g. f776 and f777).

The competition sponsor has worked to remove time-dimensionality from the data. However, the observations are still listed in order from old to new in the training set. In the test set they are in random order.