**QTM2000 – Fall 2014**

**Guidelines for the Final Team Project**

**Scope and Objective of the Project**

Some of the main objectives of this project are: improving students’ practical experience with data analytics, predictive modeling, and comparative model assessment; enhancing students’ presentation and communication skills; and accustoming students to working under uncertainty. This last objective merits further elaboration. It should be stressed, that it is possible that during the work on this project students may encounter occasional roadblocks including but not limited to data preparation, coding, implementation of known methods or models, or simply the execution time of their programs and code. Needless to say, that the less such roadblocks, the better it is for the team. At the same time, such obstacles are not uncommon when working with real engagements that involve data. Thus, should such problems arise during the project, they should be welcomed as a positive challenge deserving an effective solution. Teams are expected to exercise prudent judgment, and collaboratively overcome such challenges as a team.

The project is predictive in its nature, and thus the objective of it is either a classification or a regression (or both). Although welcomed and even encouraged during the preliminary exploratory stages of the engagement, the unsupervised learning methods (clustering or association rules) may be within the scope but should **not** be the objective of the engagement.

This project accounts for 30% of each student’s final course grade. Due to this reason, as well as for enabling the teams to exercise their best judgment and collaboratively tackle all the action items posed by the project, throughout the engagement, the instructor will have **fairly limited availability** to assist the teams with the performance of the work. Apart from the status update meetings, office hours, and occasional (quick) email communications, the instructor will not be available to intervene.

**Expectations from Teams Prior to the Presentation**

Throughout the entire project, the team members are expected to collaborate. Rather than solely focusing on individual tasks, each student is expected to take full ownership of the engagement and add her/his contribution in all stages of the work. Collaboration is in the best interest of each student, in light of the fact that each team member gets the same grade for the written deliverable report. At the completion of the project, students will be required to submit a peer-evaluation assessing the collaborative skills and effectiveness of their team members.

Finally, prior to the presentation the teams are expected to closely adhere to the items listed under the “Deadlines and Action Items” section of this document.

**Expectations from Teams During the Presentation**

The following are the core competencies that each team should demonstrate during the final presentation:

* Explain the nature and structure of the received data, the noteworthy distributions of variables, as well as any issues such as missing values, outliers, and duplicates;
* Elaborate on all the undertaken data-analytic approaches individually, as well as comparatively by effectively comparing and contrasting them;
* Emphasize and summarize the key findings as well as observations;
* Fit the presentation within the allocated timeframe;
* Deliver the presentation in a way that is well-organized and accessible to both technical and non-technical audiences; and
* Provide clear and thorough answers to all the questions asked by the instructor.

Note, that each presenter will have 5 minutes for his/her pitch except for the first presenter who will both open up the presentation for 5 minutes and conclude it for another 5 minutes. It is absolutely important that presenters time themselves prior to the final presentation. Due to the stringent 95-minute class time at the day of the presentation, the instructor may stop the presentation and move on to the next team, if a team goes beyond the allocated time.

**Expectations from Teams Regarding the Final Deliverable Report**

Listed below are some of the main characteristics that a successful deliverable report should possess:

* Discussion of the business problem/data, the objective of the project as defined by the team and confirmed by the instructor, and the appropriate data-analytic approaches undertaken by the team;
* An illustrative exploratory data analysis of the data prior to the modeling, followed by a clear explanation of the structure of the data including the distributions of variables, any issues such as potential missing values, outliers, and duplicates, as well as anything else that the team may find worthwhile to tell about the data ; and
* Careful and comprehensive analysis of the data using some of the pertinent supervised learning approaches studied during the course. The report should contain the necessary technical sophistication and depth, including but not limited to accurate model parameter estimation summaries (whenever applicable), predictive accuracy measures, comparative model assessment, etc.
* A thorough conclusion achieved as a result of the data mining performed throughout the engagement.

Note that the report should be no shorter than 10 pages and no longer than 20 pages, including all the graphs, tables, exhibits, appendixes, and bibliography (if relevant). The teams should be very selective in incorporating graphs and tables into the report, and the ones which will end up in the report should be clearly relevant to the matter, strongly related to and value-adding to the storyline.

The report should be typed in Times New Roman 12-size font using 1-inch margins on both sides of each page. In addition, page numbers and double spacing are required throughout the entire document. All figures and tables should have clear captions and legends, and should be appropriately referenced within the text. Any title, label, legend, axis, and header used in a table or a figure should be intuitive and legible. In addition, figures must have clearly labeled axes (no strange-looking variable-named titles or axes, please). If in the report references are being made to color (of graphs, displays, etc.), then the report should be printed in color and not in black-and-white. The report should be spell-checked and follow the proper rules of grammar.

**Deadlines and Action Items**

Outlined below are the major deadlines and the corresponding action items. Note that each of those deadlines is **hard**, and thus not subject to changes.

|  |  |
| --- | --- |
| **Due Date** | **Action Items** |
| Nov 3  In Class | **Project Kick-off**  The teams are announced, and the presentation date for each team is determined through a random assignment. The search for data begins. |
| Nov 5  8:00 pm | **E-mail Notification of 1st and 2nd Choice of Data**  A representative from each team emails the instructor the data set of first choice (the link, if the team suspects that the data may be too large to be emailed) that it would like to use for the project. In addition he/she also lets the instructor know **the name and the link** for the backup dataset (2nd choice) . In addition, the representative notifies the instructor of **3 possible 20-minute time slots** during which **all members** of the team will be availale to meet with the instructor for the *1st Status Update Meeting* between Monday, November 10 – Wednesday, November 12. The instructor will reply by choosing 1 of those 3 time slots. |
| Nov 10 – Nov 12 | **1st Status Update Meeting (Instructor’s Office)**  The team meets with the instructor for 20 minutes to familiarize him with the data that it has chosen to work with (just the 1st choice). In particular, the team is expected to show some exploratory data analyses of the data, and demonstrate an understanding of the strcuture of the data (size, missing values, outliers, duplicates, etc. for each variable in the data set). Any issues overlooked by the team may negatively impact the progress of the engagement, and thus it is the teams’s full responsibility to come to this meeting with a clear understanding of the data structure.  The instructor will approve the final data set within 24 hours after this meeting, and will notify each team of his decision through e-mail. |
| Nov 14  11 pm | **E-mail Notification of Roles**  A representative from each team notifies the instructor (through e-mail) the breakdown of responsibilities chosen from the roles outlined in the “Breakdown of Responsibilities” section of this document.  On this day the instructor will notify the team (through email) the exact time, during which he will be meeting the team for 20-minutes (in his office) on November 19th for the *2nd Status Update Meeting*. |
| Nov 19  Class Time | **2nd Status Update Meeting (Instructor’s Office)**  The team meets with the instructor for 20 minutes to update him on the progress since the first status update meeting, and to familiarize the instructor with the scope and the objective of the project. Topics of discussion will include – what is the problem that the team is trying to solve; is the nature of the problem calling for a classification or a prediction; what are the potential models and methods that the team has tried, and what are some of the preliminary results; etc. In addition, the team asks the questions that it has generated since the first status update meeting. |
| Nov 24-28 | **Thanksgiving Break** |
| Dec 2  6: 00 pm | **E-mail the Slides (only teams presenting on December 3)**  The teams presenting on December 3 email the instructor a **single Powerpoint file** of their presentation. While the teams are allowd to make last-minute changes to this presentation, this file is supposd to be very similar to what will be presented to the class. The teams should have the latest version of the presentation file with them on the day of the presentation. |
| Dec 3 | **Presentations of Three of the Teams** |
| Dec 7  6: 00 pm | **E-mail the Slides (only teams presenting on December 8)**  The teams presenting on December 8 email the instructor a **single Powerpoint file** of their presentation. While the teams are allowd to make last-minute changes to this presentation, this file is supposd to be very similar to what will be presented to the class. The teams should have the latest version of the presentation file with them on the day of the presentation. |
| Dec 8 | **Presentations of Rest of the Teams** |
| Dec 8  1: 00 pm | **Peer Evaluation (Qualtrics) Survey Opens**  Students gain access to the online survey. |
| Dec 10 | **Final Project Deliverable Report** **(due in class)**  **Online Peer Evaluation Closes** **(at 8:00 pm)**  **Project Concludes** |

**About Finding the Data**

The data set should contain no less than 1,000 and not more than 5,000 rows and anywhere from 10 to 50 columns. Data sets having significantly larger dimensions almost certainly will present computational issues that the teams should avoid getting into. Below are the data sources that the teams should choose their data from:

* <https://www.kaggle.com/>
* <https://archive.ics.uci.edu/ml/index.html>
* <http://www.bentley.edu/centers/csbigs/issues>
* <http://www.transtats.bts.gov/DataIndex.asp>
* <http://it.stlawu.edu/~rlock/datasurf.html>

The variables in the data should have as little missing entries as possible. The less missing data the dataset has, the easier it will be for the team to work with the data and to focus the efforts on modeling rather than on data cleansing and pre-processing. Excessive missing data are likely to adversely affect the performance of the models and methods that the team is going to develop, and thus hurt the overall project. It is very important to note, that missing data are not always left blank in the raw data files but instead may have been coded using special symbols such as 9999s, -1s, 0000s, etc. Team members must make sure that they are familiar with the coding of missing data that their data of choice uses.

Note that there exist many data sets in the above mentioned data sources that are uploaded in formats other than \*.xls or \*.csv. If that is the case with the data that the team intends to work with for this engagement, then it is the team’s full responsibility to carefully (without any loss of information) convert the data into a format readable by RStudio programs that the team will be using for this engagement. Notice also that no two teams outlined at the end of this report should use the same data for the engagement. The email notification on November 5th (see the “Deadlines and Action Items” section of this document) will be used to ensure that data sets will be approved to be used on a first-come-first-served basis. Finally, for their projects teams **cannot** choose data that has been discussed and illustrated in class.

**Breakdown of Responsibilities**

Following the engagement kick-off, the engagement team has to determine internally how to split the roles and responsibilities among its team members, and who will present what, during the final presentation. However, there are fixed responsibilities that the instructor has pre-defined for students to choose from. Note that the role that a student assumes throughout the engagement should match his/her role during the presentation. For example, the student responsible for building a logistic regression model has to present about that exact topic during the final presentation. During the initial internal team meetings, the engagement team should distribute the roles by choosing those from the second column of the corresponding table below.

|  |  |
| --- | --- |
| **Four-member Teams** | |
| **Team Member** | **Role** |
| **A** | Explain the business problem that the team is trying to solve and provide a brief background of the subject matter. Explain the structure of the data through an illustrative exploratory data analysis that may involve clustering, histograms, boxplots, and any other means that brings to light the structure of the data. Also, comment on any issues such as missing values, duplicates, or outliers (5 minutes) |
| **B** | Present the analysis of the 1st predictive model (5 minutes) |
| **C** | Present the analysis of the 2nd predictive model (5 minutes) |
| **D** | Present the analysis of the 3rd predictive model (5 minutes) |
| **A** | Compare the models 1-3 and conclude the presentation (5 minutes) |

|  |  |
| --- | --- |
| **Three-member Teams** | |
| **Team Member** | **Role** |
| **A** | Explain the business problem that the team is trying to solve and provide a brief background of the subject matter. Explain the structure of the data through an illustrative exploratory data analysis that may involve clustering, histograms, boxplots, and any other means that brings to light the structure of the data. Also, comment on any issues such as missing values, duplicates, or outliers (5 minutes) |
| **B** | Present the analysis of the 1st predictive model (5 minutes) |
| **C** | Present the analysis of the 2nd predictive model (5 minutes) |
| **A** | Compare the models 1-2 and conclude the presentation (5 minutes) |

Note that, while attending to the questions asked by the team, the instructor will **not** provide solutions but may instead direct the teams towards the right approach in case the team is significantly off track with its strategy. During the final presentation, the team is expected to defend its approach by clearly demonstrating how it works and by carefully addressing instructor’s questions.

**Grading**

Each student’s grade will be made up of **four** distinct components**:**

* Student’s individual presentation to the class on **Wednesday, December 3 or Monday, December 8.** The presentation day of each team will be determined in class based on random assignment. This component will be graded based on the overall technical accuracy and soundness of the solution, the quality of student’s answers to the questions raised by the audience (made up of the instructor and perhaps other students), as well as student’s presentational and communication skills. **This will account for 40% of student’s final project score**;
* A written deliverable report that student’s team **will submit in-class, at the beginning of the last class on Wednesday, December 10.** Each team submits a single deliverable, and each member of the team receives the same score for this component. The grading of this component will be strictly based on technical accuracy and soundness of the proposed solution, together with the clarity of exposition. **This will account for another 40% of your final project score**;
* A peer evaluation (online survey) form that each of student’s team members will submit during **Monday, December 8 – Wednesday, December 10,** assessing **student’s** effectiveness in carrying out the part of the project that student was responsible for, as well as student’s effectiveness when working with his/her team members. Note that the online survey will close at **8:00 pm on Wednesday, December 10th** and students missing this deadline will not be given another opportunity to make it up. **This will account for 10% of student’s final project score.**
* A peer evaluation (online survey), which **the student** will submit during **Monday, December 8 – Wednesday, December 10**, assessing the effectiveness of his/her team members. Note that the online survey will close at **8:00 pm on Wednesday, December 10th** and students missing this deadline will not be given another opportunity to make it up. **Note that simply submitting the completed evaluation on time will earn the student the remaining 10% of the project grade.**

**Structure of the Teams**

|  |  |  |
| --- | --- | --- |
| **Section 03** | | |
| **Team 1**  Josh Boutin  Somia Farid  Christopher Hacunda  Michael MacDonald | **Team 2**  Nina Duong  William Felix  Sebastian Gorostiaga  Rupam Gupta | **Team 3**  Nicolaas Budhiparama  Alexa Dumont  Nelson Munoz |
| **Team 4**  Tan Kabra  Hannah Kim  Yasmine Sebti  Anibha Singh | **Team 5**  James Cullinan  Andre Haddad  Arish Halani  Katherine Wood |  |

|  |  |  |
| --- | --- | --- |
| **Section 04** | | |
| **Team 1**  Abdi Abdirizak  Savannah Carlin  Chantel Jiang  Daniel Madden | **Team 2**  Jonathan Klinvex  Allana Nelson  Fuyucki Watanabe | **Team 3**  Phillip Aubrey  James Cantoni  Vu Hoang  Sergio Rodriguez |
| **Team 4**  Amy Gao  Michael Gorman  Rachel Liam  Sam Melville | **Team 5**  Mary Donahue  Samantha Donovan  David St. Peter  Kylie Stevens | **Team 6**  Jimmy Gozal  Farid Laib  Sergio Salas  Allison Windsor |