**Capstone Projects**

**Team self-assessment**

**1. After finishing your project, are you able to fulfill all objectives of your proposal?**

Please list the steps from your proposal and respective achievement rates and shortcomings. If you achieved more than what you had proposed originally, please describe your major achievements.

|  |  |  |
| --- | --- | --- |
| **Steps/items** | **% comp.** | **Comments** |
| Data extraction (ETL) | 70% | Could not load and use external data in .csv format. |
| Data preparation | 100% |  |
| Exploratory data analysis | 50% | We did not find relevant and actionable business insights. |
| Train logistic regression model | 100% |  |
| Test logistic regression model | 80% | Did not assess the model performance with ROC curve. |
| Train random forest model | 70% | We wanted to fine tune the number of predictors. |
| Test random forest model | 50% | Only test sample, did not implement OOB error. |
| Finalize results | 100% |  |
| Presentation preparation | 90% | Cosmetics issues |

**2. What are the major obstacles your team has faced to fulfill your project objectives?**

Please analyze it through several tangents.

|  |  |
| --- | --- |
| **Data collection and preparation** | It was difficult to understand the level of granularity of the different datasets and made it complicated to merge and aggregate the datasets.  We were not able to cleanly import and use the .csv files in this project but it would be a future enhancement. |
| **Business insights generation** | We were stuck in our EDA step but communication and meeting with the stakeholders were very useful to help us converge and find business insights. |
| **Proposal design** | Our initial machine learning algorithm proposal was not the most relevant choice and we eventually selected and estimated another method. |
| **Team organization and management** | Due to some external constraints, roles in the team were changed.  Some disagreements happened while preparing the development dataset: we had to confirm and validate some assumptions and definitions with the stakeholders.  Finally, we were aligned for the presentation and realized how important are others’ perspective. |
| **Technical strength/weakness** | Coding skills for data manipulations were critical to prepare a development dataset.  Data visualization and storytelling require time and brainstorming discussion to end up with relevant and actionable insights. |
| **Unforeseen issues** | We did not foresee the time that we would spend on the data preparation step and we had to make some decisions and assumptions as a team to be able to move forward.  Presentation preparation was time consuming as well. |

**3. Please list the contributions of individual team members in the projects and their roles.**

Write in % the overall contribution of each team member.

|  |  |  |
| --- | --- | --- |
| **Team members** | **% contr.** | **Comments** |
| A | 40% | A did most of the data preparation and shared EDA step with B.  They supported B to develop the random forest model. |
| B | 45% | B helped A with the data preparation and shared EDA with A. We were able to generate some business insights.  B took the lead on the random forest model development. |
| C | 15% | Due to personal constraints, C contributed less to the project and focused on the logistic regression estimation.  They contributed a lot to the final presentation though. |

**4. In finishing your project, summarize what your team, both at an individual level and at a group level, have learned in the process.**

|  |
| --- |
| We need to reinforce our coding skills in terms of data manipulation and preparation. We realized it was a very important step and is time consuming. The machine learning part did not take most of the time in the project.  It is not straightforward to find actionable and relevant insights from exploring the data.  Teamwork is important as we had different skills and preferences.  Communicating and validating the assumptions and definitions are critical to move forward in the project. |

**5. If the capstone project starts over again, what would you do differently to address the issues you have identified and encountered.**

|  |
| --- |
| I would have asked more questions to my teammates at the beginning to make sure we were aligned on the objectives and understood correctly the data.  We may think about the end product and story earlier in the process. |

**6. Can you think of other industries and topics on which you could apply the same technique and methodology you used in your capstone project?**

|  |
| --- |
| Classifier methods can be applied to any yes/no predictions or multiclass problems.  In marketing, we can anticipate customer churn  In banking, this could also be applied to fraud detection  In healthcare, we could predict health outcome (risk of cancer, ...) |

**7. Please list the names of instructors and TAs you have discussed with for your project and estimate the amount of time you have spent with them.**

|  |  |  |
| --- | --- | --- |
| **Staff members** | **Time** | **Comments** |
| Instructor 1 | 1h30 | 3 x 30 minute meetings |
| Instructor 2 | 30 | Support on random forest model selection |
| TA 1 | 2h30 | 3 x 30 minute meetings  Huge help on data preparation and model dev |
| TA 2 | 2h | Help on EDA and logistic |