**PROTOCOL TITLE: An investigation of the effect of data generation and visualization on model selection for predictive machine learning models.**

Include the full title or name of the protocol

**PRINCIPAL INVESTIGATOR (PI):**

**Name**: Remco Chang

**Department**: Computer Science

**Are you a student**? Yes  No

*If Yes, you must designate a Faculty Advisor below.*

**FACULTY ADVISOR (required for Student PIs):**

**Name**:

**Department**:

**VERSION NUMBER/DATE**

**Version 1 – 01/23/2020**

# Purpose of the study:

In this study, we will be studying the effect of various techniques of generating artificial data on the models chosen for predicting on held-out data. Traditional machine learning model selection compares aggregate measures calculated on a holdout dataset. Previous research has shown that visualizations of model predictions can help users compare models and choose a superior model than just using aggregate measures alone. This study builds on this previous research by comparing an additional technique, in which a generative Bayesian model is used to generate artificial data, demonstrate how a set of machine learning models would predict on that artificial data, and allow the user to select their model based on this artificial data.

Participants would access our system through a web interface, and be asked to select a small number of predictive machine learning models out of a set of dozens and rank them. In order to make their ranking, they will have the opportunity to have the system generate artificial data to help them make their choice. We will vary the types of data that the models are trained on for different participant groups.

The first hypothesis is that if the data used to train the models comes from a different distribution than the held-out data, then generating artificial data will help participants choose models that performs better on the held-out data.

The second hypothesis is that if the data used to train is from the same distribution as the held-out data, then we will perform no worse.

# Background / Literature Review / Rationale for the study:

Previous research has shown that visualizations of model predictions can help users compare models and choose a superior model than just using aggregate measures alone (Cashman 2019). This is in contrast to the firmly-entrenched belief of empirical risk minimization, in which a model that performs well on validation data should perform well on held-out data. Learning theory (Kearns and Vazirani 1994) tells us one explanation could be that the held-out data comes from a different distribution than the data used for training the models or calculating aggregate metrics. If this is the case, expert users may be able to extrapolate whether a given machine learning model is overfitting the distribution of the training data, or if it is a robust model that will perform well on their concept of the held-out data. The previous research showed some evidence of this, but it was unclear what led to the improved model selection. In this work, we explicitly show the user how the model would perform on out-of-sample data by generating artificial data using generative Bayesian models.

# Participant Population:

The study will be an online crowdsourcing study. Most online crowdsourcing studies in visualization are conducted by recruiting participants through Amazon Mechanical Turk, which was shown to have a participant population similar to that of typical in-person university studies by Heer & Bostock 2010. However, in this study, we plan on recruiting participants through Reddit to find participants with more analytical experience. In order to fairly assess whether the visualization helped users pick better machine learning models, we need to find a participant population that has some analytical knowledge, knows the dataset well, and has incentive to generate a good predictive model.

We plan to post the experiment to the online message board reddit.com/r/baseball, a message board which typically has around 5000 concurrent users at any given time. The members of this message board are often very interested in analytics because of the statistical revolution of the last 10-15 years in baseball. The experiment will ask participants to choose the model that will most effectively predict whether a player will be inducted into the hall of fame, a common discussion topic among baseball fans. They will likely have intimate knowledge of the dataset, which is a publicly-available list of players and their career stats. The recruitment text is attached as a separate document.

A study by Jamnik and Lane in 2017 demonstrated that data gathered from participants recruited from Reddit and participants recruited in the university setting were statistically similar.

# Special Populations:

|  |  |
| --- | --- |
|  | Children |
|  | Fetuses/Neonates |
|  | Prisoners |
|  | Members of the military |
|  | Non-English speakers |
|  | Those unable to read (illiterate) |
|  | Employees of the researcher |
|  | Students of the researcher |
|  | Adults lacking capacity to consent and/or adults with diminished capacity to consent, including, but not limited to, those with acute medical conditions, psychiatric disorders, neurologic disorders, developmental disorders, and behavioral disorders |
|  | Disadvantaged in the distribution of social goods and services such as income, housing, or healthcare |
|  | Fear of negative consequences for not participating in the research (e.g. institutionalization, deportation, disclosure of stigmatizing behavior) |
|  | Approached for participation in research during a stressful situation such as emergency room setting, childbirth (labor), etc. |

# Research Locations and Sample Size:

* 1. **Research Locations**

The research will be conducted entirely online through a web interface.

***5.2* Sample Size**

The difference between means for power analysis in this case is the difference between a model’s performance that is overfit on the training data and a model that fits well on the training data. This is likely a small difference. For the sake of power analysis, we believe the difference could be about 10% - the accuracy is 80% vs. 70%. Using the sample size calculator at <https://www.stat.ubc.ca/~rollin/stats/ssize/n2.html>, and assuming a standard deviation of 10%, we get a sample size of *n=16* participants. Since there are two conditions, we will need double that, so *n=32* participants. We believe this is feasible if gathered from our expected participant population. If we are lucky enough to get more participants signed up, we will use all participant data in our analysis. This will enable us to do additional analysis on subgroups of the population, such as participants who report prior experience with machine learning or visualization.

# Procedures Involved:

We have developed a visual analytics application that enables a user to select from a set of pre-trained machine learning models. We will start with open source baseball data that is publicly available online. This data will consist of the career batting numbers of baseball players in Major League Baseball, along with a label of whether they made the hall of fame or not. The data will be split into a training dataset, a validation dataset, and a held-out dataset for evaluation. A number of machine learning models will be trained using the training dataset, and participants will be asked to select from a set of these machine learning models.

Participants will be recruited through the message board reddit.com/r/baseball. If they click the link to participate in the study, the participants will go through the following process in the experiment. We expect the process will take 30-60 minutes.

* Electronic consent form
* Demographic survey
* Survey on their experience with analytics, machine learning, data science, and visualization (see attached).
* Then, they will be brought to our application. In our application (see attached screenshots), they will explore a dataset of career numbers for Major League Baseball Players. They will be able to inspect a set of machine learning models’ hall of fame predictions on that dataset. They will be asked about a number of fictional players and their predictions on whether that player would be elected to the hall of fame, and then based on how the models would predict that player, they will select and rank the top 5 of the models presented to them.
* Lastly, they’ll be brought to a page thanking them for participating in the experiment and explaining the purpose of the experiment.

There are two conditions that we will be testing, differing in what data the machine learning models are trained and what data the users see. These two conditions will be compared against a baseline condition.

* The baseline condition represents traditional model selection, in which the top five models selected would be the top five models by accuracy.
* The first condition, seen by half of the participants, is whether they will choose more accurate models on heldout data than the baseline if both the participants and the baseline only see models trained on a "bad" set of data (just second basemen).
* The second condition tested, seen by the other half of the participants, will mirror the first condition, except the models will be trained on the full distribution of players.

We will report a one-tailed T test between zero and the difference in accuracy of the best model selected by the participant and the baseline model's accuracy for both conditions. We anticipate that when the training dataset is only 2nd baseman, this value will be significantly positive, and when the training dataset is the full dataset, this test will confirm the null hypothesis. We will also report precision at top k, a metric commonly used in machine learning for ranking tasks that measures the precision of the top 5 models chosen (by both baseline and participant) against the true top performing 5 models for the held out data. We will also analyze whether experience with analytics, machine learning, data science, or visualization have effect on these metrics, using Chi-Squared.

In addition to recording the top 5 models from each participant, we will record interactions with the system to make a qualitative analysis to understand the efficacy of various features in the tool.

Because the study will be shared online, we want to be wary of participants sharing answers and tactics. To get around this, we make effort to have the user experience of each participant be slightly different. This means that rather than only train the two dozen models that a single participant will use, we will train around 100 models and subsample from that set for each participant. In addition, we will also subsample the validation set that the users see so that they also have different sets of data points to interact with. This will not effect our data collection or analysis. During data collection, we will also take efforts to prevent the same participant from participating multiple times. To do this, we will store a cookie in the participant's browser's local storage that will expire one month after they first participate. This cookie will be checked when the page first loads, and if the participant has already participated, they will see a message explaining that they are only allowed to participate one time. This information will only live on the participant's browser and we will not use cookies to keep track of any other information.

In order to generate artificial data, we will use a generative bayesian model with a gaussian process prior. This method will choose points that are most disagreed upon by the list of models, i.e. it will choose data points that maximize the shannon entropy of the set of predictions of the models viewed by the user.

* 1. **Additional Safeguards for Special Populations**:   
     N/A

# Investigational Medical Devices:

N/A

# Incomplete Disclosure or Deception:

N/A

# Recruitment Methods:

We plan to post the experiment to the online message board reddit.com/r/baseball, a message board which typically has around 5000 concurrent users at any given time. The members of this message board are often very interested in analytics because of the statistical revolution of the last 10-15 years in baseball. If we plan on updating the IRB with additional sources for participants. We believe data collection will only take around a week. Recruitment text is attached.

# Consent Process:

Consent will be obtained through the web via a consent form (see attached) that potential participants will view when they first click the link to our study given out in the recruitment text. There are no foreseeable risks to participating. Participants are told in both the recruitment script and the consent text that they must be 18 or over, located in the United States, and fluent in English. This is a standard method for receiving consent for experiments that take place on the web.

# Compensation:

No compensation will be provided to participants. Participants will not be responsible for any costs because of participation in research.

# Economic Burden:

Subjects will not be under any economic burden due to participating in research.

# Recording with Audio, Video, or Photographs

No audio, video, or photographs will be recorded.

# Potential Benefits to Participants:

Participants may benefit by viewing the baseball data in our system. It may help them understand machine learning models. It may help them understand the dataset better.

# Risks to Participants:

There are no foreseeable risks to participants.

# Withdrawal of Participants:

If participants exit the application before completing the task, their data will be discarded during the analysis.

# Data Management and Confidentiality:

The web application will send data that is gathered to a secure server run in the computer science department of Tufts University. It will only be accessible to the investigators of the study as outlined on this IRB application. It will be stored for the minimum of 3 years. Participants won’t be re-contacted for any reason.

The data collected will consist of the demographic survey data, the survey of experience with experience with analytics, machine learning, data science, and visualizations. It will also include the list of machine learning models selected by the participants and the baseline for their experiment. In addition, we will record the interactions of the participants with the system, as logged by our application. This will include which elements the user clicks on and the timing of their clicks.

If a participant does not consent, their corresponding data will be discarded. If they exit the application before completing the task, their corresponding data will be discarded.

# Provisions to Protect the Privacy and Confidentiality of Participants and the Research Data:

No personally identified information will be recorded per participant. There is no risk of privacy or confidentiality being in danger.

All participants will be walked through a set of interactions before they have control of the system via a tutorial they click through. The tutorial will be accessible at any point during the experiment.

# Provisions to Monitor the Data to Ensure the Safety of Subjects:

Research involves only minimal risk to subjects.

# Compensation for Research-Related Injury:

Research involves only minimal risk to subjects.

# Data Sharing and Specimen Banking:

Data will not be shared outside of the investigators listed on this protocol.

# International Research:

All investigators reside in the United States, and all participants are required to reside in the United States.

# Multiple sites:

All investigators are affiliated only with Tufts University.

# Reliance Agreements/Single IRB:

N/A

# Qualifications to Conduct Research and Resources Available:

Dr. Remco Chang has been the PI for several IRB approved studies at Tufts University. Dr. Chang and Dylan Cashman recently conducted IRB study #1902023 entitled “User-Guided Neural Architecture Search for Discovering Image Classifiers” last March.