CS 6250 Homework 1

Tian Tan

Jan.25, 2017

1 CITI Certification

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

COMPLETION REPORT - PART 1 OF 2
COURSEWORK REQUIREMENTS*

* NOTE: Scores on this Requirements Report reflect quiz completions at the time all requirements for the course were met. See list below for details. See separate Transcript Report for more recent quiz scores, including those on optional (supplemental) course elements.

- **Name:** Tian Tan (ID: 6056537)
- **Email:** ttan40@gatech.edu
- **Institution Affiliation:** Georgia Institute of Technology (ID: 324)

- **Curriculum Group:** Human Research
- **Course Learner Group:** Group 1 Biomedical research Investigators and Key Personnel
- **Stage:** Stage 1 - Basic Course

- **Report ID:** 21946253
- **Completion Date:** 15-Jan-2017
- **Expiration Date:** 15-Jan-2020
- **Minimum Passing:** 70
- **Reported Score*:** 88

REQUIRED AND ELECTIVE MODULES ONLY	DATE COMPLETED	SCORE
Belmont Report and CITI Course Introduction (ID: 1127)	15-Jan-2017	3/3 (100%)
History and Ethics of Human Subjects Research (ID: 498)	15-Jan-2017	6/7 (86%)
Basic Institutional Review Board (IRB) Regulations and Review Process (ID: 2)	15-Jan-2017	4/5 (80%)
Informed Consent (ID: 3)	15-Jan-2017	5/5 (100%)
Social and Behavioral Research (SBR) for Biomedical Researchers (ID: 4)	15-Jan-2017	1/4 (25%)
Records-Based Research (ID: 5)	15-Jan-2017	3/3 (100%)
Genetic Research in Human Populations (ID: 6)	15-Jan-2017	5/5 (100%)
Populations in Research Requiring Additional Considerations and/or Protections (ID: 16680)	15-Jan-2017	5/5 (100%)
Vulnerable Subjects - Research Involving Children (ID: 9)	15-Jan-2017	3/3 (100%)
Vulnerable Subjects - Research Involving Pregnant Women, Human Fetuses, and Neonates (ID: 10)	15-Jan-2017	2/3 (67%)
International Studies (ID: 971)	15-Jan-2017	3/3 (100%)
FDA-Regulated Research (ID: 12)	15-Jan-2017	5/5 (100%)
Research and HIPAA Privacy Protections (ID: 14)	15-Jan-2017	4/5 (80%)
Vulnerable Subjects - Research Involving Workers/Employees (ID: 483)	15-Jan-2017	4/4 (100%)
Conflicts of Interest in Research Involving Human Subjects (ID: 488)	15-Jan-2017	5/5 (100%)
Avoiding Group Harms - U.S. Research Perspectives (ID: 14080)	15-Jan-2017	3/3 (100%)
Stem Cell Research Oversight (Part I) (ID: 13882)	15-Jan-2017	3/5 (60%)

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: www.citiprogram.org/verify/?k6f931bdf-32d2-4b9d-9573-51c5b07a98a5-21946253

CITI Program
Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

COMPLETION REPORT - PART 2 OF 2
COURSEWORK TRANSCRIPT**

** NOTE: Scores on this Transcript Report reflect the most current quiz completions, including quizzes on optional (supplemental) elements of the course. See list below for details. See separate Requirements Report for the reported scores at the time all requirements for the course were met.

- **Name:** Tian Tan (ID: 6056537)
- **Email:** ttan40@gatech.edu
- **Institution Affiliation:** Georgia Institute of Technology (ID: 324)

- **Curriculum Group:** Human Research
- **Course Learner Group:** Group 1 Biomedical research Investigators and Key Personnel
- **Stage:** Stage 1 - Basic Course

- **Report ID:** 21946253
- **Report Date:** 15-Jan-2017
- **Current Score**:** 88

REQUIRED, ELECTIVE, AND SUPPLEMENTAL MODULES	MOST RECENT	SCORE
History and Ethics of Human Subjects Research (ID: 498)	15-Jan-2017	6/7 (86%)
Informed Consent (ID: 3)	15-Jan-2017	5/5 (100%)
Social and Behavioral Research (SBR) for Biomedical Researchers (ID: 4)	15-Jan-2017	1/4 (25%)
Belmont Report and CITI Course Introduction (ID: 1127)	15-Jan-2017	3/3 (100%)
Records-Based Research (ID: 5)	15-Jan-2017	3/3 (100%)
Genetic Research in Human Populations (ID: 6)	15-Jan-2017	5/5 (100%)
Vulnerable Subjects - Research Involving Children (ID: 9)	15-Jan-2017	3/3 (100%)
Vulnerable Subjects - Research Involving Pregnant Women, Human Fetuses, and Neonates (ID: 10)	15-Jan-2017	2/3 (67%)
FDA-Regulated Research (ID: 12)	15-Jan-2017	5/5 (100%)
International Studies (ID: 971)	15-Jan-2017	3/3 (100%)
Research and HIPAA Privacy Protections (ID: 14)	15-Jan-2017	4/5 (80%)
Vulnerable Subjects - Research Involving Workers/Employees (ID: 483)	15-Jan-2017	4/4 (100%)
Conflicts of Interest in Research Involving Human Subjects (ID: 488)	15-Jan-2017	5/5 (100%)
Avoiding Group Harms - U.S. Research Perspectives (ID: 14080)	15-Jan-2017	3/3 (100%)
Basic Institutional Review Board (IRB) Regulations and Review Process (ID: 2)	15-Jan-2017	4/5 (80%)
Stem Cell Research Oversight (Part I) (ID: 13882)	15-Jan-2017	3/5 (60%)
Populations in Research Requiring Additional Considerations and/or Protections (ID: 16680)	15-Jan-2017	5/5 (100%)

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: www.citiprogram.org/verify/?k6f931bdf-32d2-4b9d-9573-51c5b07a98a5-21946253

Collaborative Institutional Training Initiative (CITI Program)
Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

COMPLETION REPORT - PART 1 OF 2
COURSEWORK REQUIREMENTS*

* NOTE: Scores on this Requirements Report reflect quiz completions at the time all requirements for the course were met. See list below for details. See separate Transcript Report for more recent quiz scores, including those on optional (supplemental) course elements.

- **Name:** Tian Tan (ID: 6056537)
- **Email:** ttan40@gatech.edu
- **Institution Affiliation:** Georgia Institute of Technology (ID: 324)

- **Curriculum Group:** Human Research
- **Course Learner Group:** Group 2 Social / Behavioral Research Investigators and Key Personnel
- **Stage:** Stage 1 - Basic Course

- **Report ID:** 21947690
- **Completion Date:** 15-Jan-2017
- **Expiration Date:** 15-Jan-2020
- **Minimum Passing:** 70
- **Reported Score*:** 96

REQUIRED AND ELECTIVE MODULES ONLY	DATE COMPLETED	SCORE
Belmont Report and CITI Course Introduction (ID: 1127)	15-Jan-2017	3/3 (100%)
Students in Research (ID: 1321)	15-Jan-2017	5/5 (100%)
History and Ethical Principles - SBE (ID: 490)	15-Jan-2017	5/5 (100%)
Defining Research with Human Subjects - SBE (ID: 491)	15-Jan-2017	5/5 (100%)
The Federal Regulations - SBE (ID: 502)	15-Jan-2017	5/5 (100%)
Assessing Risk - SBE (ID: 503)	15-Jan-2017	4/5 (80%)
Informed Consent - SBE (ID: 504)	15-Jan-2017	5/5 (100%)
Privacy and Confidentiality - SBE (ID: 505)	15-Jan-2017	5/5 (100%)
Research with Children - SBE (ID: 507)	15-Jan-2017	4/5 (80%)
Research in Public Elementary and Secondary Schools - SBE (ID: 508)	15-Jan-2017	5/5 (100%)
International Research - SBE (ID: 509)	15-Jan-2017	5/5 (100%)
International Studies (ID: 971)	15-Jan-2017	3/3 (100%)
Internet-Based Research - SBE (ID: 510)	15-Jan-2017	5/5 (100%)
Research and HIPAA Privacy Protections (ID: 14)	15-Jan-2017	4/5 (80%)
Vulnerable Subjects - Research Involving Workers/Employees (ID: 483)	15-Jan-2017	4/4 (100%)
Conflicts of Interest in Research Involving Human Subjects (ID: 488)	15-Jan-2017	5/5 (100%)

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: www.citiprogram.org/verify/?k8c507a68-2ba1-4e44-a9c9-2fb10e4e0fe1-21947690

CITI Program
Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

COMPLETION REPORT - PART 2 OF 2
COURSEWORK TRANSCRIPT**

** NOTE: Scores on this Transcript Report reflect the most current quiz completions, including quizzes on optional (supplemental) elements of the course. See list below for details. See separate Requirements Report for the reported scores at the time all requirements for the course were met.

- **Name:** Tian Tan (ID: 6056537)
- **Email:** ttan40@gatech.edu
- **Institution Affiliation:** Georgia Institute of Technology (ID: 324)

- **Curriculum Group:** Human Research
- **Course Learner Group:** Group 2 Social / Behavioral Research Investigators and Key Personnel
- **Stage:** Stage 1 - Basic Course

- **Report ID:** 21947690
- **Report Date:** 15-Jan-2017
- **Current Score**:** 96

REQUIRED, ELECTIVE, AND SUPPLEMENTAL MODULES	MOST RECENT	SCORE
Students in Research (ID: 1321)	15-Jan-2017	5/5 (100%)
History and Ethical Principles - SBE (ID: 490)	15-Jan-2017	5/5 (100%)
Defining Research with Human Subjects - SBE (ID: 491)	15-Jan-2017	5/5 (100%)
Belmont Report and CITI Course Introduction (ID: 1127)	15-Jan-2017	3/3 (100%)
The Federal Regulations - SBE (ID: 502)	15-Jan-2017	5/5 (100%)
Assessing Risk - SBE (ID: 503)	15-Jan-2017	4/5 (80%)
Informed Consent - SBE (ID: 504)	15-Jan-2017	5/5 (100%)
Privacy and Confidentiality - SBE (ID: 505)	15-Jan-2017	5/5 (100%)
Research with Children - SBE (ID: 507)	15-Jan-2017	4/5 (80%)
Research in Public Elementary and Secondary Schools - SBE (ID: 508)	15-Jan-2017	5/5 (100%)
International Research - SBE (ID: 509)	15-Jan-2017	5/5 (100%)
International Studies (ID: 971)	15-Jan-2017	3/3 (100%)
Internet-Based Research - SBE (ID: 510)	15-Jan-2017	5/5 (100%)
Research and HIPAA Privacy Protections (ID: 14)	15-Jan-2017	4/5 (80%)
Vulnerable Subjects - Research Involving Workers/Employees (ID: 483)	15-Jan-2017	4/4 (100%)
Conflicts of Interest in Research Involving Human Subjects (ID: 488)	15-Jan-2017	5/5 (100%)

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: www.citiprogram.org/verify/?k8c507a68-2ba1-4e44-a9c9-2fb10e4e0fe1-21947690

Collaborative Institutional Training Initiative (CITI Program)
Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

COMPLETION REPORT - PART 1 OF 2
COURSEWORK REQUIREMENTS*

* NOTE: Scores on this Requirements Report reflect quiz completions at the time all requirements for the course were met. See list below for details. See separate Transcript Report for more recent quiz scores, including those on optional (supplemental) course elements.

- **Name:** Tian Tan (ID: 6056537)
- **Email:** ttan40@gatech.edu
- **Institution Affiliation:** Georgia Institute of Technology (ID: 324)

- **Curriculum Group:** CITI Health Information Privacy and Security (HIPS)
- **Course Learner Group:** CITI Health Information Privacy and Security (HIPS) for Biomedical Research Investigators
- **Stage:** Stage 1 - HIPS

- **Report ID:** 21946254
- **Completion Date:** 15-Jan-2017
- **Expiration Date:** 15-Jan-2020
- **Minimum Passing:** 70
- **Reported Score*:** 92

REQUIRED AND ELECTIVE MODULES ONLY	DATE COMPLETED	SCORE
Basics of Health Privacy (ID: 1417)	15-Jan-2017	15/16 (94%)
Health Privacy Issues for Researchers (ID: 1419)	15-Jan-2017	3/5 (60%)
Basics of Information Security, Part 1 (ID: 1423)	15-Jan-2017	5/5 (100%)
Basics of Information Security, Part 2 (ID: 1424)	15-Jan-2017	5/5 (100%)
Protecting Your Computer (ID: 1425)	15-Jan-2017	8/8 (100%)

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: www.citiprogram.org/verify/?k3901a11a-1fbd-4a0c-a991-cd591e1117dd-21946254

CITI Program
Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

COMPLETION REPORT - PART 2 OF 2
COURSEWORK TRANSCRIPT**

** NOTE: Scores on this Transcript Report reflect the most current quiz completions, including quizzes on optional (supplemental) elements of the course. See list below for details. See separate Requirements Report for the reported scores at the time all requirements for the course were met.

- **Name:** Tian Tan (ID: 6056537)
- **Email:** ttan40@gatech.edu
- **Institution Affiliation:** Georgia Institute of Technology (ID: 324)

- **Curriculum Group:** CITI Health Information Privacy and Security (HIPS)
- **Course Learner Group:** CITI Health Information Privacy and Security (HIPS) for Biomedical Research Investigators
- **Stage:** Stage 1 - HIPS

- **Report ID:** 21946254
- **Report Date:** 15-Jan-2017
- **Current Score**:** 92

REQUIRED, ELECTIVE, AND SUPPLEMENTAL MODULES	MOST RECENT	SCORE
Basics of Health Privacy (ID: 1417)	15-Jan-2017	15/16 (94%)
Health Privacy Issues for Researchers (ID: 1419)	15-Jan-2017	3/5 (60%)
Basics of Information Security, Part 1 (ID: 1423)	15-Jan-2017	5/5 (100%)
Basics of Information Security, Part 2 (ID: 1424)	15-Jan-2017	5/5 (100%)
Protecting Your Computer (ID: 1425)	15-Jan-2017	8/8 (100%)

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: www.citiprogram.org/verify/?k3901a11a-1fbd-4a0c-a991-cd591e1117dd-21946254

Collaborative Institutional Training Initiative (CITI Program)
Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org

2 Descriptive Statistics

Metric	Deceased patients	Alive patients	Function to complete
Event Count			event_count_matrics
1. Average Event Count	982.014	498.118	
2. Max Event Count	8635	12627	
3. Min Event Count	1	1	
Encounter Count			encounter_count_matrics
1. Average Encounter Count	23.038	15.452	
2. Max Encounter Count	203	391	
3. Min Encounter Count	1	1	
Record Length			record_length_matrics
1. Average Encounter Count	127.532	159.2	
2. Max Encounter Count	1972	2914	
3. Min Encounter Count	0	0	

Table 1: Descriptive statistics for alive and dead patients

3 Feature Construction

4 Predictive Modeling

4.1 Model Creation

Model	Accuracy	AUC	Precision	Recall	F-Score
Logistic Regression	0.9545	0.9454	0.9869	0.8988	0.9408
SVM	0.9940	0.9945	0.9882	0.9970	0.9926
Decision Tree	0.7763	0.7476	0.7922	0.6012	0.6836

Table 2: Model performance on training data

Model	Accuracy	AUC	Precision	Recall	F-Score
Logistic Regression	0.7381	0.7375	0.6804	0.7333	0.7059
SVM	0.7381	0.7389	0.6768	0.7444	0.7090
Decision Tree	0.6714	0.6569	0.6330	0.5555	0.5917

Table 3: Model performance on test data

Strategies:
1. Gather more data (if retriving data is not expensive) since taking more data reduces random errors by the law of large numbers.
2. Feature engineering to generate more information from existing data. New generated features maybe explains variance of the training data better, e.g. create new features such as a percentage of events number versus total duration from the first event date to last date.
3. Running several algorithms and compare their performance, tune parameters by grid search to find the best parameters.
4. Ensemble different models by bagging or boosting, since combine weak models might produce stronger ones.
5. Cross validation to achieve more generalized relationships and reduce variability.

4.2 Model Validation

CV strategy	Accuracy	AUC
K-Fold	0.7213	0.7075
Randomized	0.7357	0.7188

Table 4: Cross Validation

4.3 Self Model Creation

Methods: I've tried three models: XGBClassifier, RandomForestClassifier and ExtraTreeClassifier, and use grid search and cross validation(10 folds) to find each model's best parameters. I choose ExtraTreeClassifier model as my final model, since it has the highest AUC value and grid score. The reason I choose these three models is because they are well-developed, frequently used supervised learning algorithms that generally have good performance on classificaiton. As for performance, my predictive model has higher auc value than all previous models. Moreover, my model has higher recall value and lower precision value, i.e. my model detects more positive cases but also get more false alarms, which suggests that my model might not doing as well as previous ones (logistic regression and SVM) with respect to accuracy, but is better in performance than Decision Tree models.