Mini Project 1:

Failure Data Analysis for Software-as-a-Service (SaaS) Business Application

ECE 313

Prof. Ravi K. Iyer

TA: Homa Alemzadeh

First yet Realistic Empirical Analysis

- You are an analyst working at the company ACME
- Your duties consist in quantifying how much a specific system deployed by ACME is reliable
  - i.e., if it fails, how frequent it fails and how it fails

- You use data extracted from computer logs collected at runtime by the system under analysis
  - Considered computer logs contain information whether a given operation performed by the considered system succeeded or failed at time T
  - The logs are collected in the text format
  - Example (from your computers): event logs (windows), syslogs (linux), Osx Logs (Mac Osx)

Description of the Datasets

- Failure data from a production cloud computing Datacenter
- The computer processes files uploaded by the user
- The processing steps in several computing stages and it can fail according to two failure types “USER DATA FAILURE”, “PLATFORM FAILURE”
- The data is structured in the following fields:

Submission Time	Computing Stage in the Failure	Computation Start Time	Computation End Time	Failure Cause	Failure Details	Failure Type
7/1/12 0:02	IT1	7/1/12 0:02	7/1/12 0:02	File Not Received	Went over cut-off time	USER DATA FAILURE
7/1/12 0:58	IT4_L2	7/1/12 0:59	7/1/12 0:59	System Error	Package Validation/Execution Failed.	PLATFORM_FAILURE
7/1/12 0:59	IT4_L2	7/1/12 0:59	7/1/12 0:59	System Error	Package Validation/Execution Failed.	PLATFORM_FAILURE
7/1/12 1:01	IT4_L2	7/1/12 1:01	7/1/12 1:02	System Error	Package Validation/Execution Failed.	PLATFORM_FAILURE
7/1/12 1:01	IT4_L2	7/1/12 1:01	7/1/12 1:01	System Error	Package Validation/Execution Failed.	PLATFORM_FAILURE

Project TASK no. 0

Get familiar with the analysis environment

1. Download and Install R:
   - http://www.r-project.org
2. Explore the R software environment and its basic commands by reading the provided tutorials on the class website

3. Import the dataset assigned to your group in R
   - **Hint:** Text files are in standard Unix format (UTF-8)
4. Split the data into two different datasets
   - **Hint:** in the provided datasets, entries with a field "USER DATA FAILURE" or “PLATFORM_FAILURE”, belong to two different type of events

Project TASK no. 1

What are the system failure modes/types?

1. What is the probability that a file submission fails?
   - **Hint:** Check out the System and Data Description document to find out the sample space.
2. Calculate the probability that there is one of the following possible failure types
   - User Data Failure, Platform Failure
     - **Hint:** Calculate the probabilities by relative frequency.
3. For each failure type, calculate the probability that the failure is due to each specific **failure cause** (present in the data).
   - E.g., what is the probability of a platform failure due to “File not found”, “Timeout error”, or etc?

Project TASK no. 1

What are the system failure modes/types?

4. Find the probability of each failure type by adding the partial failure probabilities in 3.
   - **Hint:** If x% of the failures are platform failures, of which 92% of them are due to the “timeout error”, 8% “file not found”, then failures due to “timeout errors” account for 92% * x% of the total failures in the whole space of events
   - How else would you arrive to this value?

Project TASK no. 2

Failure rate: How often does the system fail?

1. Compute the failure rate on daily basis (failures per day) for:
   1. All failure events taken together
   2. For each of the failure types in Task 1.2

2. Visualization of the results
   A. Draw the scatter plot of the data points against time
      - **Hint:** X axis is time, Y axis is the computed failure rate values
   B. Draw a histogram for the computed failure rates.

   – All the charts should be appropriately formatted by showing the legend, axis title, and chart title.
      - **Hint:** R provides a great online HELP. Try invoking: *help(NAME_OF_THE_COMMAND)* from the R shell
   – Provide a short and intuitive description of the results
      - **Hint:** A few bullets on the two/three most important things you notice

Project TASK no. 3

Calculate the time between two consecutive failures

- For each failure type in Task 1
  - Add a column to the dataset computing the time between two consecutive failure entries
    - Plot the results
    - Plot the the average time between two consecutive failures per week for the whole period
    - Provide a short and intuitive description of the results (2-3 bullets)
  - **Bonus**: compute the probability that two consecutive failures (of the same type) happen within 1 minute, 5 minute, 10 minute, 15 minutes….until 30 minutes from each other and plot the results
    - Question: How would you explain the results intuitively?
      - Explain in 3 bullets

Project Timeline and grading

- **Checkout the class website for description of system, tasks, and data sets**: http://courses.engr.illinois.edu/ece313/SectionB/projects.html
- **Your group members and data sets are assigned to you in the class and by Email.**
- **A report (including both the R code and results)** must be delivered electronically to the TA: **Email to ece313.B@gmail.com.**
  - Explain all your work and include the code with comments
  - Non-readable reports will be returned without a grade.
  - Write your names, group name, and the data-set assigned to you on the report.
- **Post your questions to ECE 313 web-board on the my.ece website.**

- **Task 0-1: due on Tuesday, September 3, 11:59 PM.**
- **Task 2-3: due on Friday, September 6, 11:59 PM.**
- **Grading**
  - **Task 0 – 1.1: 25%**
  - **Task 1.2 – 1.3: 30%**
  - **Task 2.1 – 2.2: 20%**
  - **Task 3: 25%**
  - **Bonus: 5%**