Project Plan
Email Classification using Machine Learning

The Capstone Experience

Team Accenture
Sofia Colella
Varsha Odapally
Griffin Carr
Kevin Wilson
Yuyu Su

Department of Computer Science and Engineering
Michigan State University
Fall 2019
Functional Specifications

• Problem: There are about 15 billion spam and phishing emails per day.
• Solution: Create a web application dashboard for analysts
• Solution: Create and enhance classification and clustering models to help triage incoming emails, such as malicious attachment, emails with URL that have dangerous payloads, emails that lead to credential phishing and spam
Design Specifications

• Results of ML algorithm will be displayed on the user’s dashboard
  ▪ Overall statistics and examples of malicious emails
• Three types of users: strategic, operational and tactical
• Strategic users see the overall results in the form of pie charts and a heat map
• Operational users are given more information regarding the malicious emails in addition to pie charts
• Tactical users are given the most information. They also have the option to take a course of action for each individual malicious email
Screen Mockup: Strategic User Dashboard
Screen Mockup: Operational User Dashboard
Screen Mockup: Tactical User Dashboard
## Screen Mockup: See More Page

### Malicious Attack Identification

- **Filtered by:** Threat Level

<table>
<thead>
<tr>
<th>Threat Level</th>
<th>Datetime</th>
<th>Location</th>
<th>Sender</th>
<th>Subject</th>
<th>Industry</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>04/03/2019 03:23:39</td>
<td>East Lansing, MI</td>
<td>PHISHER A</td>
<td>Hey There!</td>
<td>Health</td>
<td>850</td>
</tr>
<tr>
<td>3</td>
<td>09/02/2019 13:24:05</td>
<td>Austin, TX</td>
<td>PHISHER B</td>
<td>Help!</td>
<td>Business</td>
<td>1547</td>
</tr>
<tr>
<td>2</td>
<td>06/23/2019 04:09:42</td>
<td>Los Angeles, CA</td>
<td>PHISHER C</td>
<td>Click now!</td>
<td>Marketing</td>
<td>487</td>
</tr>
<tr>
<td>1</td>
<td>08/27/2018 11:13:50</td>
<td>Chicago, IL</td>
<td>PHISHER D</td>
<td>Winner!</td>
<td>Medical</td>
<td>589</td>
</tr>
<tr>
<td>1</td>
<td>03/22/2017 12:40:50</td>
<td>New York, NY</td>
<td>PHISHER E</td>
<td>DANGER!</td>
<td>Automotive</td>
<td>400</td>
</tr>
<tr>
<td>1</td>
<td>04/03/2019 03:23:39</td>
<td>Freemont, MI</td>
<td>PHISHER F</td>
<td>Click this!</td>
<td>Education</td>
<td>347</td>
</tr>
<tr>
<td>1</td>
<td>12/24/2018 08:09:33</td>
<td>Seattle, WA</td>
<td>PHISHER G</td>
<td>Expired insur.</td>
<td>Entertainment</td>
<td>80</td>
</tr>
<tr>
<td>1</td>
<td>01/03/2016 11:11:04</td>
<td>Boston, MA</td>
<td>PHISHER H</td>
<td>Verify credit</td>
<td>Sports</td>
<td>572</td>
</tr>
<tr>
<td>1</td>
<td>02/06/2017 03:10:00</td>
<td>Portland, OR</td>
<td>PHISHER J</td>
<td>URGENT</td>
<td>Fashion</td>
<td>233</td>
</tr>
</tbody>
</table>

1–50 of 995
Technical Specifications

• Web app, database, machine learning model hosted on CentOS VM
• Machine learning back-end
• Python API call back-end
• MongoDB on VM with parsed email data, MongoDB on AWS hosted cluster for login info
• Flask front-end to display dashboards
System Architecture
System Components

• Software Platforms / Technologies
  ▪ Python 3.7
  ▪ Tensorflow 1.14.0
  ▪ AWS
  ▪ MongoDB
  ▪ VirtualBox VM - CentOS
  ▪ Flask
  ▪ PyCharm
Risks

• Ranking Harmful Emails (High)
  ▪ Determining a ranking system to list emails in terms of threat level
  ▪ Threats will be evaluated using the “Threat Triangle.” Also known in cyber threat intelligences as capability, intent, and opportunity.

• Output Interpretation (High)
  ▪ Model outputs run the risk of being misinterpreted, based on misunderstanding/incorrect assumption of how the desired model was built.
  ▪ Clear idea of how model must be built, what assumptions must be made, and what the output should tell the audience.

• Under/Overfitting Model (Medium)
  ▪ Having “High Bias” (underfitting) or “High Variance” (overfitting) which can lead to poor predictions for future used data sets.
  ▪ Train-Test Split of our data (Train model on 70% of data, measure error rate on remaining 30% of data)

• Data (low)
  ▪ Poor data quality, lack of data, lack of variability of data
  ▪ Client provided us mock up data
Questions?