שלום חברים,
מצ"ב הטיפים שלי לעבודה מתודית שתוביל להצלחה בתחרות:

1. ידע מקצועי **(domain knowlege)** :
דברו עם אנשי הסייבר של הקבוצה ונסו להבין כל אחד מהמשתנים של הדאטסט. מה המשמעות של המשתנה, למה משמש והאם יש לו קשר כלשהו לסוג כרטיס הרשת על המחשב הנייד של התוקף.
2. אנליזה של הנתונים **(EDA)**:
בדקו את ההתפלגות של כל משתנה, תספרו את מספר הערכים השונים שיש לכל משתנה. בדקו אם יש חסרים ותנסו להבין את מנגנון היצירה שלהם. בדקו את החסרים גרפית עם heatmap.
בדקו האם יש ערכי קיצון ואיך ניתן לטפל בהם. בדקו קורלציות בין המשתנים השונים. בדקו האם יש הבדלים המשתנים השונים לבין סוג כרטיס הרשת. בדקו האם יש כפילות של שורות. תעזרו באנשי ב-**BI** בקבוצה שלכם. שתפו את אנשי הסייבר מהממצעים של ה-EDA.
3. נקו את נתונים ואשרו אותם עם משתנים חדשים (feature extraction / feature engineering).
4. בחרו את המשתנים הכי רלוונטיים **(feature selection)**.
השתמשו בכמה שיטות כדי שיהיה לכם בחירה יותר איכותית.
5. חלקו את נתונים ובדקו שדאטסטים מאוזנים.
זכרו להשאיר דאטסט לסוף התהליך (TEST) ,כדי שתוכלו לבדוק את המודל הסופי לפני העלאת התוצאה ל-KAGGLE.
6. בחרו את המטריקה שלפי דעתכם הכי נכון להשתמש (מטריקה אחת בלבד).
7. בחירת המודל **(model selection)**:
אמנו לפחות 5 מודלים שונים ובחרו את הטוב מבניהם. חשוב להשוות בין התוצאות של ה-TRAIN וה-DEV.
8. הדקו את המודל הסופי (**fine tuning**)
9. בדקו את טיב המודל כולו ולפי כל סוג של יצרן כרטיס רשת בנפרד.
בדקו מהם המשתנים המשפיעים ביותר. נסו להבין מה הסיבה שמקרים שהשתייכו לקבוצה הלא נכונה (misclassification) לא היו נכונים. שלב זה ייתן לכם אינדיקציות לבעיות עם הנתונים. את השלב הזה מומלץ לעשות עם כל הצוות.
10. במידת הצורך, בצעו שינויים לנתונים כדי לנסות לשפר את המודלים.
11. שמרו את כל המודלים והעבודה שעשיתם.
אל תכתבו על הקוד הקיים ואל תמחקו קוד.
12. את תהליך הניקוי והוספת משתנים חדשים (3) כתבו בצורה שתאפשר להשתמש באותו קוד עבור נתונים חדשים (קובץ ה-TEST שתקבלו ב-KAGGLE כדי לייצר את הפרדיקציה שלכם).

מאחל לכם הצלחה רבה!
סומך עליכם שלא תאפשרו להאקר לפרוץ למערכות האוניברסיטה ,להצפין ולדרוש כופר.
ד"ר תומס קרפטי