SABANCI ÜNİVERSİTESİ YÖNETİCİ GELİŞTİRME BİRİMİ

İLERİ VERİ ANALİTİĞİ AKADEMİSİ

Sabancı Üniversitesi

İLERİ VERİ ANALİTİĞİ AKADEMİSİ HACKATHON

January 5-6, 2023

SABANCI

PROBLEM

- **Predicting car insurance claims**
  - You are a start-up company which **provides insurance for cars**.
  - Your company managed to acquire important number of customers, and now wants to optimize the cost of the insurance by a**nalyzing data and drawing insights.**
  - You are asked to **build a machine learning model for** predicting whether a policyholder will make a claim in the next 6 months.

Features

Target Variable:

- **is claim**: either 0 or 1 (it will be 1 if a policyholder will make a claim in the next 6 months, 0 otherwise)

Regular Features:

- ncap rating: Safety rating given by NCAP (out of 5)
- is power door locks: Boolean flag indicating whether a power door lock is available in the car or not.
- policyholder age: Normalized age of policyholder in years
- is parking camera: Boolean flag indicating whether the parking camera is present in the car or not.
- rear brakes type: Type of brakes used in the rear of the car
- is adjustable steering: Boolean flag indicating whether the steering wheel of the car is adjustable or not.
- is tpms: Boolean flag indicating whether Tyre Pressure Monitoring System is present in the car or not.

Features

Regular Features:

- is driver seat height adjustable: Boolean flag indicating whether the height of the driver seat is adjustable or not.
- segment: Segment of the car (A/ B1/ B2/ C1/ C2).
- car age: Normalized age of the car in years.
- is central locking: Boolean flag indicating whether the central locking feature is available in the car or not.
- is rear window wiper: Boolean flag indicating whether the rear window wiper is available in the car or not.
- height: Height of the car (Millimetre).
- cluster area: Area cluster of the policyholder.
- is ecw: Boolean flag indicating whether Engine Check Warning is available in the car or not.
- fuel type: Type of fuel used by the car.
- torque: Maximum Torque generated by the car.
- engine volume: Engine displacement/engine capacity/engine size of the car.

Features

Regular Features:

- transmission type: Transmission type of the car.
- manufacturer: Encoded manufacturer/company of the car.
- cylinder: Number of cylinders present in the engine of the car.
- is rear window washer: Boolean flag indicating whether the rear window washer is available in the car or not.
- is front fog lights: Boolean flag indicating whether front fog lights are available in the car or not.
- is brake assist: Boolean flag indicating whether the brake assistance feature is available in the car or not.
- is power steering: Boolean flag indicating whether power steering is available in the car or not.
- is esc: Boolean flag indicating whether Electronic Stability Control is present in the car or not.
- population: Population density of the city (Policyholder City).
- is rear window defogger: Boolean flag indicating whether rear window defogger is available in the car or not.

Features

Regular Features:

- time period: Time period of the policy.
- engine type: Type of engine used in the car.
- is speed alert: Boolean flag indicating whether the speed alert system is available in the car or not.
- steering type: Type of the power steering present in the car.
- length: Length of the car (Millimetre).
- width: Width of the car (Millimetre).
- is parking sensors: Boolean flag indicating whether parking sensors are present in the car or not.
- power: Maximum Power generated by the car (bhp@rpm).
- gross weight: The maximum allowable weight of the fully-loaded car, including passengers, cargo and equipment (Kg).

Features

Regular Features:

- is day night rear view mirror: Boolean flag indicating whether day & night rearview mirror is present in the car or not.
- model: Encoded name of the car.
- gear box: Number of gears in the car.
- airbags: Number of airbags installed in the car.
- turning radius: The space a vehicle needs to make a certain turn (Meters).
- area danger level: Danger level of area by starting 1 upto 5

PROBLEM

- **Pros**
  - Familiar domain ☺️
  - Real-life & trendy data
  - Medium sized dataset (43,592 samples and 43 features)

- **Challenges:**
  - There are 2 tasks!
  - The data has missing values, categorical variables, …

Task Overview

Task Overview

- **You are expected to complete 2 tasks:**
  - **Task 1 (50pts):** In the **visualization and insight task**, you are not told what to look for. You need to analyze the data and draw **insights that may be used in the business.**
    - Ex: **At what car age is there a significant increase/drop in possibility of a claim?**
  - **Task 2 (50pts):** In the **prediction task** (organized as a Kaggle competition), you are expected to build a machine learning model for the given task

- **The presentation format given in the next slides will also outline what is expected from you.**

Scoring

OVERALL SCORE

TASK 1 –Data Visualization & Insights – 50pts

- **Presentation Quality – 25pts**
  - **Think that you have a short time to pitch your idea to the company**
  - Good presentation skills: right level of detail, interesting, readable slides, good timing,…
  - Good visuals
  - Good answers to questions

- **Value of Insights – 25pts**
  - Importance will be measured as **business** or **information value**.
    - Could this information be predicted anyway (e.g. “The cars with higher ncap score are safer that the ones with lower ncap score”) or is it surprising/informative?
    - Can you suggest ways to derive business value out of your insights?
      - The company can use the information you provide to find a new way to get the attraction of the customers.

OVERALL SCORE

- **TASK 2: Kaggle task – 50pts**
  - **Approach – 25pts**
    - You will be given the maximum (25) points as default; but flaws or shortcomings in your approach will result in points (2-5pts) subtracted from this (e.g. not trying enough suitable approaches (e.g. at least 3 separate approaches), applying an unsuitable transformation, wrong approach in model selection, incorrect handling of the missing values etc.).
    - Note that even though your approach will also affect your test performance of your system, the two are graded separately so that your score does not depend exclusively on your test performance.
  - **Test Performance – 25pts**
    - **Your system will be ranked/graded according to AUC (Area Under the Receiver Operating Characteristic Curve) score**
    - Non-working systems or groups who don't submit any test results will get 0 points from this category.
    - Best performing system will obtain the maximum score (25)
    - The rest of the systems will be graded linearly between 0-25, according to their performance.

Presentation

TASKS and PRESENTATION

- Each presentation should be a maximum of **11** slides altogether.
- **Duration:** You will be given **7 minutes in total** to present your work and a separate 2-3 minutes to answer questions from the jury.
  - We recommend you spend
    - **3-4 minutes** on visualization/insights into the problem (Task 1)
    - **3-4 minutes** on technical details for the prediction task (Task 2).
  - Groups who do not respect the time limits will lose 5pts.
- **Presenters**: 2 or 3 people from each group are expected to present (at least one data scientist/engineer and one business translator).
- **Title slide** of your presentation should contain a 1-line project code-name (describing the project) and your group number and members.

TASKS and PRESENTATION

- **Task 1:** 4-5 slides should contain **visualizations** that will summarize important aspects of the data and and **insights** that may suggest business value/potential.
  - If you have multiple insights, it would be good to present them in a coherent fashion (grouped together etc) and it would be good if there is some business value of the insights
  - Order your insights (most important first)
  - Graphics should be readable (maybe max 4 graphics on a slide)
  - **Task 1 should be presented by Business Translators**

TASKS and PRESENTATION

- **Task 2:** 4-5 slides should contain **your approach to the prediction part of the problem.**
  - **You must include:**
    - Preprocessing if any.
    - Feature selection/extraction if any. Which features emerged as most important and if you have found any **new** features that are useful
    - **The methods/algorithms tried and the one that resulted in your best cross-validation performance and corresponding results.**
    - Analysis of your system performance – what can you say about when/how system is working well or failing? Any critical issues, challenges or solution steps in your approach.
  - **Task 2 should be presented by Data Scientists/Engineers.**

PRESENTATION

- The **last slide** should contain a 1- or 2-line summary of **the IVAA and Hackathon experience** by each of the group members.
  - This is an opportunity where you can reflect on
    - how much you have learned
    - what you have liked/disliked in the program
    - how you may transfer the knowledge and experience acquired to solve problems at your job.

Details

IMPLEMENTATION

- **Programming language**: **You are expected to use a Python 3 notebook.**
- **Approach and Software**:
  - You may make use of *any* available toolboxes and libraries commercial or free.
  - You may use *any* approach (any classifier including those that were not addressed during the training program; and any data normalization; dimensionality reduction, feature extraction techniques, etc.).
  - Make sure your preprocessing/feature extraction can be done easily on the test data; you will need to apply the same feature processing steps on the test data (Pipelines, ColumnTransformers, etc **HIGHLY RECOMMENDED**).
- **Computing Resources**: You may use *any* computing resources available, including the cloud and the HPC.

KAGGLE DETAILS

- Here is the link for Kaggle.
- train.xlsx
  - 43,592 instances, 43 regular features (except ID column), and 1 target feature (*is claim*)
- test.csv
  - 15000 instances, 43 regular features (except ID column)

- Remember that the performance metric is AUC (Area Under Curve)
- We have Public (30%) / Private (70%) Leaderboard (their weighted average will be considered) → Test_score = 0.3 * Public_score + 0.7 * Private_score

KAGGLE SUBMISSION FILE

- It should be **csv file**
- There should be header
  - ***ID,Probability***
- "ID"s should be an integer 0 to 14999 (both included)
- "Probability"s should be a real numeric value between 0 and 1 (both included).
- Check ***sample_submission.csv*** in Kaggle link
- Groups can make at most 7 submissions in a day (don't forget to compose your groups in Kaggle with name "GroupX" like Group1, Group2, etc.)
  - Participants will need to wait until the next UTC day after submitting the maximum number of daily submissions.

CODE SUBMISSION

- **You should submit your code that produces the predictions as .ipynb** file. Please use this form to submit your ipynb and data files.
  - No need to submit code written to generate visualization or to draw insights.
  - You need to include your input data as you might have used external data etc

- **Your submitted code must give the exact test results when trained with the given training data using your code.**
  - In order to remove any randomness, please start by setting a **random number seed (or random_state)** to have results that you can replicate.

Rules

RULES/DISQUALIFICATION

- You may **use** information and even codes found on the internet with proper credit. However;
  - **All code and work should be your own (taking only “snippets” of code is OK).**
  - **Your team should be able to explain any part of your submitted code.**

- Disqualification:
  - Failure to give proper credit or in general any scientifically unethical behaviour or
  - Using outside data will result in disqualification of your team,

  will result in disqualification.

CRITICAL DEADLINES

- **January 6th 10:45** Test Results & Code Delivery & Kaggle Deadline
- **January 6th 12:15** Presentation Delivery
- **January 6th 13:30** Team Presentations Start

Questions?