《数据挖掘与商务智能》课程 考核项目

# 项目要求：

## 项目编码要求

结合你所抽选的题目,采用python进行数据挖掘模型的构建，小组上交的编码必须包括以下五个部分。

Module 1: Importing Data Sets

Module 2: Data Wrangling

Module 3: Exploratory Data Analysis

Module 4: Model Development

Module 5: Model Evaluation and Refinement

## 项目汇报要求

要求每组提交PPT，阐述以下几个方面：

1. 项目任务
2. 数据源分析
3. 数据预处理
4. 数据探索性分析
5. 建模过程，可以采用几种模型来分析
6. 评价，对不同模型进行评估

## 成绩考核

* 编码是否完成，运行情况
* PPT制作的美观度
* PPT内容的完整度
* 汇报语言的清晰流畅
* 报告（word）——每人分工

# 项目案例以及要求

## Heart disease failure

### Task:

Create a model to assess the likelihood of a death by heart failure event. This can be used to help hospitals in assessing the severity of patients with cardiovascular diseases.

### Data:

Dataset from Davide Chicco, Giuseppe Jurman: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20, 16 (2020)

Datasource: heart\_failure\_clinical\_records\_dataset

## Health Insurance

### Task:

Your client is an Insurance company that has provided Health Insurance to its customers now they need your help in building a model to predict whether the policyholders (customers) from past year will also be interested in Vehicle Insurance provided by the company.

Just like medical insurance, there is vehicle insurance where every year customer needs to pay a premium of certain amount to insurance provider company so that in case of unfortunate accident by the vehicle, the insurance provider company will provide a compensation (called ‘sum assured’) to the customer.

Building a model to predict whether a customer would be interested in Vehicle Insurance is extremely helpful for the company because it can then accordingly plan its communication strategy to reach out to those customers and optimise its business model and revenue.

Now, in order to predict, whether the customer would be interested in Vehicle insurance, you have information about demographics (gender, age, region code type), Vehicles (Vehicle Age, Damage), Policy (Premium, sourcing channel) etc.

Evaluation Metric

The evaluation metric for this hackathon is ROC\_AUC score.

Public and Private split

The public leaderboard is based on 40% of test data, while final rank would be decided on remaining 60% of test data (which is private leaderboard)

### data:

Our client is an Insurance company that has provided Health Insurance to its customers now they need your help in building a model to predict whether the policyholders (customers) from past year will also be interested in Vehicle Insurance provided by the company.

An insurance policy is an arrangement by which a company undertakes to provide a guarantee of compensation for specified loss, damage, illness, or death in return for the payment of a specified premium. A premium is a sum of money that the customer needs to pay regularly to an insurance company for this guarantee.

For example, you may pay a premium of Rs. 5000 each year for a health insurance cover of Rs. 200,000/- so that if, God forbid, you fall ill and need to be hospitalised in that year, the insurance provider company will bear the cost of hospitalisation etc. for upto Rs. 200,000. Now if you are wondering how can company bear such high hospitalisation cost when it charges a premium of only Rs. 5000/-, that is where the concept of probabilities comes in picture. For example, like you, there may be 100 customers who would be paying a premium of Rs. 5000 every year, but only a few of them (say 2-3) would get hospitalised that year and not everyone. This way everyone shares the risk of everyone else.

Just like medical insurance, there is vehicle insurance where every year customer needs to pay a premium of certain amount to insurance provider company so that in case of unfortunate accident by the vehicle, the insurance provider company will provide a compensation (called ‘sum assured’) to the customer.

Building a model to predict whether a customer would be interested in Vehicle Insurance is extremely helpful for the company because it can then accordingly plan its communication strategy to reach out to those customers and optimise its business model and revenue.

Now, in order to predict, whether the customer would be interested in Vehicle insurance, you have information about demographics (gender, age, region code type), Vehicles (Vehicle Age, Damage), Policy (Premium, sourcing channel) etc.

Data source: Health Insurance

## HR Analytics: Job Change of Data Scientists

### Task Details

This dataset designed to understand the factors that lead a person will work for the company(leaving current job) ,and the goal of this task is building model(s) that uses the current credentials, demographics, experience to predict the probability of a candidate looking for a new job or will work for the company.

The whole data divided to train and test . Also sample submission has been provided correspond to enrollee\_ id of test set ( enrolle\_ id | target)

* The dataset is imbalanced so it might affect your result if you dont handle it
* Most features are categorical (Nominal, Ordinal, Binary), some with high cardinality so encoding methods and techniques will help to boost models performance
* Missing imputation strategy might affect the results so it can be a part of your pipeline as well.

### Data

A company which is active in Big Data and Data Science wants to hire data scientists among people who successfully pass some courses which conduct by the company. Many people signup for their training. Company wants to know which of these candidates are really wants to work for the company after training or looking for a new employment because **it helps to reduce the cost and time as well as the quality of training or planning the courses and categorization of candidates**. Information related to demographics, education, experience are in hands from candidates signup and enrollment.

This dataset designed to understand the factors that lead a person to leave current job for HR research too. By model(s) that uses the current credentials, demographics, experience data you will predict the probability of a candidate to look for a new job or will work for the company, as well as interpreting affected factors on employee decision.

The whole data divided to train and test . Target isn't included in test but the test target values data file is in hands for related tasks. A sample submission correspond to enrollee\_id of test set provided too with columns : enrollee \_id , target

**Features**

* enrollee\_id : Unique ID for candidate
* city: City code
* city\_ development \_index : Developement index of the city (scaled)
* gender: Gender of candidate
* relevent\_experience: Relevant experience of candidate
* enrolled\_university: Type of University course enrolled if any
* education\_level: Education level of candidate
* major\_discipline :Education major discipline of candidate
* experience: Candidate total experience in years
* company\_size: No of employees in current employer's company
* company\_type : Type of current employer
* lastnewjob: Difference in years between previous job and current job
* training\_hours: training hours completed
* target: 0 – Not looking for job change, 1 – Looking for a job change
* Predict the probability of a candidate will work for the company
* Interpret model(s) such a way that illustrate which features affect candidate decision

Data source: Job Change of Data Scientists

## Credit Card customers

### Improve Performance of predicting churned customers.

Our top priority in this business problem is to identify customers who are getting churned. Even if we predict non-churning customers as churned, it won't harm our business. But predicting churning customers as Non-churning will do. So recall (TP/TP+FN) need to be higher. Till now, I have managed to get a recall of 62%. Need better.

### Data

A manager at the bank is disturbed with more and more customers leaving their credit card services. They would really appreciate if one could predict for them who is gonna get churned so they can proactively go to the customer to provide them better services and turn customers' decisions in the opposite direction

I got this dataset from a website with the URL as https://leaps.analyttica.com/home. I have been using this for a while to get datasets and accordingly work on them to produce fruitful results. The site explains how to solve a particular business problem.

Now, this dataset consists of 10,000 customers mentioning their age, salary, marital\_status, credit card limit, credit card category, etc. There are nearly 18 features.

We have only 16.07% of customers who have churned. Thus, it's a bit difficult to train our model to predict churning customers.

Data source: BankChurners.csv

## Real Estate Investment Trust

### Project

In this assignment, you are a Data Analyst working at a Real Estate Investment Trust. The Trust will like to start investing in Residential real estate. You are tasked with determining the market price of a house given a set of features. You will analyze and predict housing prices using attributes or features such as square footage, number of bedrooms, number of floors, and so on. A template notebook is provided in the lab; your job is to complete the ten questions. Some hints to the questions are given in the template notebook.

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.

### Data

**id** : A notation for a house

**date**: Date house was sold

**price**: Price is prediction target

**bedrooms**: Number of bedrooms

**bathrooms**: Number of bathrooms

**sqft\_living**: Square footage of the home

**sqft\_lot**: Square footage of the lot

**floors** :Total floors (levels) in house

**waterfront** :House which has a view to a waterfront

**view**: Has been viewed

**condition** :How good the condition is overall

**grade**: overall grade given to the housing unit, based on King County grading system

**sqft\_above** : Square footage of house apart from basement

**sqft\_basement**: Square footage of the basement

**yr\_built** : Built Year

**yr\_renovated** : Year when house was renovated

**zipcode**: Zip code

**lat**: Latitude coordinate

**long**: Longitude coordinate

**sqft\_living15** : Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area

**sqft\_lot15** : LotSize area in 2015(implies-- some renovations)

data source: kc\_house\_data\_NaN.csv

## Loan train

### project

In this notebook we try to practice all the classification algorithms that we learned in this course.

We load a dataset using Pandas library, and apply the following algorithms, and find the best one for this specific dataset by accuracy evaluation methods.

Now, it is your turn, use the training set to build an accurate model. Then use the test set to report the accuracy of the model You should use the following algorithm:

* K Nearest Neighbor(KNN)
* Decision Tree
* Support Vector Machine
* Logistic Regression

\*\* Notice:\*\*

You can go the pre-processing, feature selection, feature-extraction, and so on, to make a better model. You should use either scikit-learn, Scipy or Numpy libraries for developing the classification algorithms.

### Data:

This dataset is about past loans. The **Loan\_train.csv** data set includes details of 346 customers whose loan are already paid off or defaulted. It includes following fields:

| **Field** | **Description** |
| --- | --- |
| Loan\_status | Whether a loan is paid off on in collection |
| Principal | Basic principal loan amount at the |
| Terms | Origination terms which can be weekly (7 days), biweekly, and monthly payoff schedule |
| Effective\_date | When the loan got originated and took effects |
| Due\_date | Since it’s one-time payoff schedule, each loan has one single due date |
| Age | Age of applicant |
| Education | Education of applicant |
| Gender | The gender of applicant |

Datasource: loan train.csv

## Heart disease

We have a data which classified if patients have heart disease or not according to features in it. We will try to use this data to create a model which tries predict if a patient has this disease or not.

Data: heartdisease.csv

1. age:代表患者年龄，以年为单位。

2. sex:代表患者性别，1为男性，0为女性。

3. cp:胸痛类型，具体类型未说明

4. trestbps：入院前的静息血压，单位为毫米汞柱

5. chol:患者血清胆固醇含量，单位mg/dl

6. fbs:患者空腹血糖大于120mg/dl，1为是，0为否

7. restecg：静态心电图结果

8. thalach:达到最大的心率值

9. exang：运动诱发心绞痛，1为是，0为否。

10. oldspeak：相对于休息来说运动引起的ST段抑制

11. slope: 运动引起的 ST段最高值斜率

12. ca: 通过荧光检测技术显示出来的主要血管数量（0-3）

13. thal:1为正常，2为固定缺陷，3为可逆缺陷。

14. target：0或1为是否患病