2-3 datasets per person w/ potential business problem and supervised + unsupervised ML model

Anusha -

1. Bankruptcy data on American companies, can use to predict if companies will go bankrupt. Use KMeans classification and logistic regression <https://www.kaggle.com/datasets/utkarshx27/american-companies-bankruptcy-prediction-dataset>
2. Predicting music rating
3. <https://www.kaggle.com/datasets/solomonameh/spotify-music-dataset>
4. <https://www.kaggle.com/datasets/samiraalipour/genomics-of-drug-sensitivity-in-cancer-gdsc>

Devi -

Mahima -

1. [E-commerce shipping data](https://www.kaggle.com/datasets/prachi13/customer-analytics/data)

Business problem: Help a COO reduce delayed deliveries and improve logistics efficiency

Supervised + unsupervised ML model: Clustering (delivery risk segmentation + classification (on-time vs. delayed deliveries)

1. [HR analytics data](https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset/code)

Business problem: Help a Chief Human Resources Officer predict employee attrition and implement retention strategies

Supervised + unsupervised ML model: Clustering (employee type segmentation) + classification models (employee attrition vs no attrition)

Nidhi -

1. [Loan Approval data](https://www.kaggle.com/datasets/taweilo/loan-approval-classification-data)

Business problem: Help reduce manual work by automating loan approvals for low-risk applicants, and minimize financial losses by identifying high-risk applicants likely to default.

1. [Weather data](https://www.kaggle.com/datasets/nikhil7280/weather-type-classification/data)

Business problem: Unpredictable weather conditions, such as heavy rain or snow, can cause delays and increase costs. Help reduce delays and improve safety,by avoiding areas potentially impacted by adverse weather conditions.

Canva link: <https://www.canva.com/design/DAGgJbiDAi4/BuKkwTvN8e2_R3QDn3f2_g/edit?utm_content=DAGgJbiDAi4&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton>

Timelines:

EDA done by 2/21

Modeling by 3/4

Presentation 3/8

Data structure and pre-processing:

Explaining the dataset, no NAs, proportion of response variable

Encoding

EDA graph - Anusha

Distribution of attrition by job roles

Distribution of attribution by marital status

Box plot of monthly income by attrition

Box plot of job satisfaction by attrition

Correlation matrix

Drop - Devi

JobLevel, YearsAtCompany, YearswithCurrManager, EmployeeNumber, StandardHours, EmployeeCount, Over18

Clustering:

Part 1: Separate Unsupervised + supervised learning

K-means with dummy coding categorical variables and other numerical variables - Devi

K-prototypes with mixed data types - Mahima

Compare the two and pick best one with most distinct clusters ^

Then do logistic regression on each cluster with cross-validation - Anusha

Random forest and gradient boosting with cross-validation - Nidhi

Part 2: Hybrid Unsupervised + supervised learning

Do cluster-wise logistic regression (k-means or k-prototype based on which one is better/feasibility)

Mixed data types clustering guide: <https://bpostance.github.io/posts/clustering-mixed-data/>

K-prototypes vs k-means guide: <https://antonsruberts.github.io/kproto-audience/>

**Next steps:**

Devi - drop columns, make 80-20 split for data and upload, kmeans

Mahima - kprototypes

Mahima and Devi finalize clusters by the weekend so Anusha and Nidhi can start second part

Anusha - finish data slides, EDA slides (with slides color coding), logistic regression on clusters w cross validation

Nidhi - random forest on clusters and gradient boosting with cross validation

Next meeting: Tuesday 3/4 before class

Deliverables:

* Mahima - fix kprototypes, slide for kprototypes cluster explanation, work on cluster-wise regression
* Devi - coalesce kmeans to the main file, kmeans vs kprototypes why you’re picking kprototypes (explain what each model does, the scores are better for kprototypes), HW part 2
* Anusha - add confusion matrices, do slides about logistic regression, work on cluster-wise regression
* Nidhi - add confusion matrices, do slides about decision trees, HW part 2

Next meeting: Friday 3/7

Need to coalesce EDA files and add descriptions/insights in python file

**OPTION A**

1. **Identify a key stakeholder (e.g., Chief Marketing Officer, CEO, Chief Information officer, Chief Finance Officer, etc.) for whom you are doing this project.** - Chief Human Resources Officer
2. **Describe the goals of the analysis?** – predict employee attrition and implement retention strategies
3. Develop an “Analysis” plan including details of
   1. Goals of analysis
   2. Analytical Plan
   3. Sampling methods that were used for data collection
   4. Data description
   5. Expected results at the end of the project.
4. The analysis should use a combination of unsupervised and supervised methods that will be discussed in the class such as – cluster analysis, factor analysis, cluster-wise regressions, tree models, multinomial logit, discriminant analysis, etc.
5. Solutions should be validated using Hold-out sampling methodology or Cross-validation methodology.
6. If you decide to use another methodology, you must write a methodology section describing it – make sure you discuss with me the methodology you plan to choose and why.
7. All the analytical commands that you use – from reading data into R, cleaning and performing hygiene on it, performing univariate summaries, multivariate modeling, etc. other analysis, etc. should be provided.
8. Write a report at the end of the project – that has an executive summary at the beginning - with strategic implications for the key stakeholders. Remember – the summary of analysis is the key to the project. The success of the project depends on satisfaction of the key stakeholders.
9. Plan to make a video presentation of your project, to be submitted the last week of the class.

Remember: an ideal project answers the question: “What insight did the project provide the key stakeholders”? What value did it provide? In other words, if these analyses were not conducted, what value would the stakeholder not have?