Week08 – GLM Investigation

Module/ Framework/Package	Name and description of the algorithm	An example of a situation where using the provided GLM implementation provides superior performance compared to that of base R or its equivalent in Python (identify the equivalent in Python
Base R	It uses Iteratively Reweighted Least squares for parameter estimation in GLMs. IRLS is an extension of weighted least squares adapted for maximum least squares adapted for maximum likelihood estimation of GLM coefficients.	It is suitable for small to medium datasets, with built-in support for various link functions and family distributions. However, due to memory constraints, it struggles with large datasets(primary limitation). Equivalent in python: **statsmodels** GLM
Big Data R	It implements parallelized GLM fitting using packages like biglm or speedglm. These methods optimize memory usage and parallel processing capabilities for large datasets	It is ideal for datasets exceeding memory limits on a single machine. Equivalent in Python: dask-ml or pyspark.ml
Dask ML	It uses an approximate, incremental solver for GLMs, optimizing with parallel and out-of-core computations. Supports L1 and L2 regularization for robust modeling.	It excels in handling distributed datasets across clusters. It outperforms base R for large-scale logistic regression tasks. Equivalent in R: biglm
Spark R	It uses a distributed IRLS-based optimization technique for GLMs, utilizing Spark's in-memory computation to handle big data efficiently.	It is effective for massive datasets stored in distributed systems like Hadoop. It is equivalent in python: pyspark.ml
Spark optimization	It utilizes gradient-based methods like Limited-	It is superior to Base R when dealing with highly scattered

	memory BFGS (L-BFGS) and Stochastic Gradient Descent (SGD) for optimizing GLMs in a distributed computing environment	datasets or when high-performance parallel computation is required. Equivalent in python : pyspark.ml or dask-ml
Scikit - learn	It Implements GLMs using coordinate descent (Lasso), and gradient descent. It also efficiently supports L1, L2, and elastic net regularization.	It outperforms Base R in scenarios requiring regularization or handling multicollinearity. Equivalent in R: glmnet