Homework 2: Cross-Validation & Clustering: Credit Approval (kNN/SVM) + Iris (k-means)

Georgia Institute of Technology, Business Analytics

Introduction to Analytics Modeling

Professor Joel Sokol

September 3, 2025

Files submitted: homework2_answers.pdf (this doc), homework2.Rmd

Question 3.1

Using the same data set (credit_card_data.txt or credit_card_data-headers.txt) as in Question 2.2, use the ksvm or kknn function to find a good classifier:

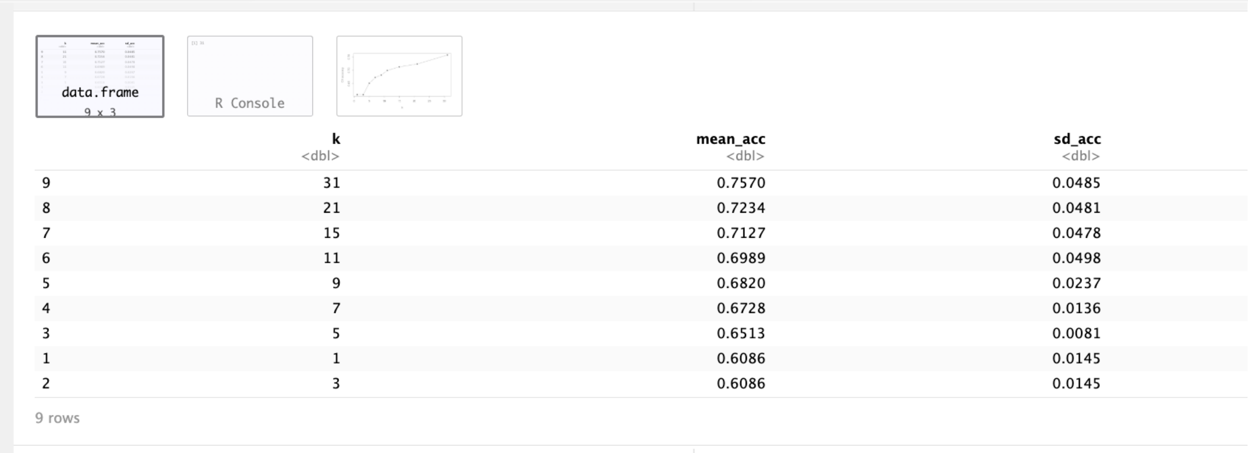
(a) **using cross-validation (do this for the k-nearest-neighbors model; SVM is optional); and**

Methodology - kNN with Cross-Validation

- **Data:** credit_card_data-headers.txt with 10 numeric predictors (A1, A2, A3, A8, A9, A10, A11, A12, A14, A15). Response variable: R1 (binary).
- **Preprocessing:** Standardized features during training (kNN sensitive to scale). Standardization applied fold-by-fold using training set statistics to avoid leakage.
- **Cross-Validation:** Stratified 5-fold CV with preserved class balance.
- **Parameter tuning:** Swept k ∈ {1, 3, 5, 7, 9, 11, 15, 21, 31}. For each k, recorded fold accuracy mean ± sd.
- **Reproducibility:** set.seed(42)

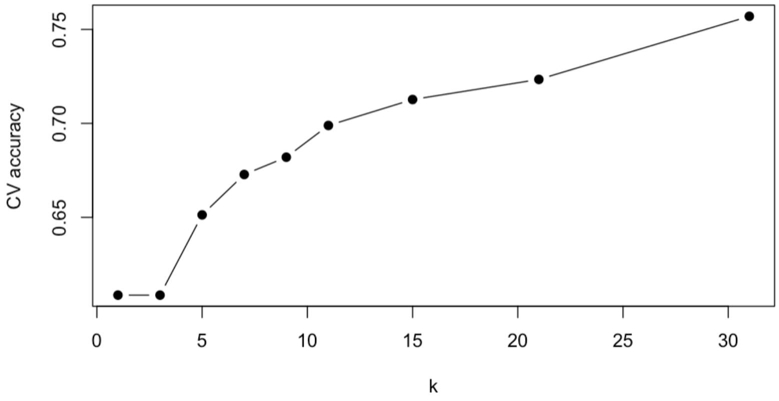
Results - kNN with Cross-Validation

- **Best k: 31**
- **CV accuracy: 0.7600 ± 0.0462**
- Accuracy increased monotonically with k and plateaued near 31 neighbors.



	k	mean_acc	sd_acc
	<dbl>	<dbl>	<dbl>
9	31	0.7570	0.0485
8	21	0.7234	0.0481
7	15	0.7127	0.0478
6	11	0.6989	0.0498
5	9	0.6820	0.0237
4	7	0.6728	0.0136
3	5	0.6513	0.0081
1	1	0.6086	0.0145
2	3	0.6086	0.0145

9 rows



Discussion - kNN with Cross-Validation

Accuracy improves steadily as k increases, plateauing near k=31, suggesting variance reduction benefits outweigh added bias. Performance exceeds the majority baseline (358/654 ≈ 0.547) by ~21 percentage points, showing meaningful predictive signal. Future extensions could test distance-weighted kernels or a wider k grid.

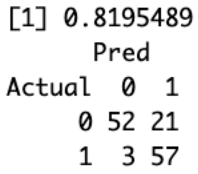
(b) splitting the data into training, validation, and test data sets (pick either KNN or SVM; the other is optional).

Methodology - SVM

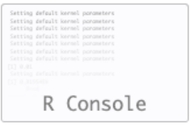
- **Data split:** Stratified 60/20/20 split (set.seed(99)) → Train (393 obs), Validation (131), Test (130).
- **Model:** Linear SVM via ksvm(kernel="vanilladot", scaled=TRUE).
- **Parameter tuning:** C ∈ {0.001, 0.01, 0.1, 1, 10, 100, 1000}. Best C chosen by validation accuracy.
- **Final model:** Retrained on Train+Val with C*. Evaluated once on Test set.

Results - SVM

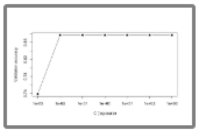
- **Selected C*: 0.01**
- **Test accuracy: 0.8195**
- **Test confusion matrix:**

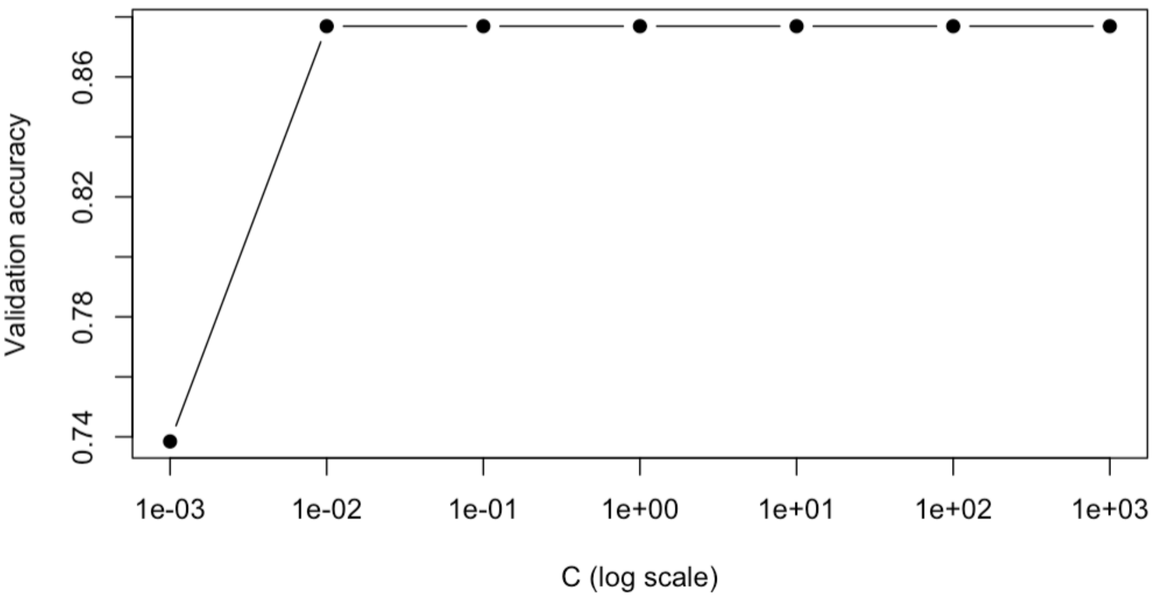


```
[1] 0.8195489
      Pred
Actual  0  1
     0 52 21
     1  3 57
```



```
Setting default kernel parameters
Setting default kernel parameters
Setting default kernel parameters
Setting default kernel parameters
Setting default kernel parameters
Setting default kernel parameters
Setting default kernel parameters
[1] 0.01
Setting default kernel parameters
[1] 0.8195489
```





Derived metrics (class “1” = positive): **precision 0.950**, **recall 0.731**, **balanced accuracy 0.838**

Discussion - SVM

The linear SVM is **conservative about approvals** (very few false positives: 3), trading recall for precision on class 1 (21 false negatives). This may be preferable when the cost of a false approval is higher than that of a missed approval; otherwise, threshold tuning or class-weighted SVMs could rebalance the trade-off. Note that kNN’s **CV accuracy** (0.76) and SVM’s **held-out test accuracy** (0.82) use different evaluation schemes and are not directly comparable.

Question 4.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a clustering model would be appropriate. List some (up to 5) predictors that you might use.

Situation. Segment users by listening behavior to improve personalized recommendations and campaigns.

Why clustering. No ground-truth “taste” labels exist; unsupervised clustering (e.g., k-means on standardized features) is appropriate to uncover natural groups.

Predictors (≤5).

1. Average daily listening minutes
2. Genre mix (% by genre)
3. Time-of-day distribution
4. Device usage mix (mobile/desktop/speaker)
5. New-music discovery rate (% new tracks last 30 days)

Expected outcome. Distinct segments (e.g., late-night indie, all-day pop, weekend-jazz) to drive playlist curation, promotions, and model personalization.

Question 4.2

The *iris* data set iris.txt contains 150 data points, each with four predictor variables and one categorical response. The predictors are the width and length of the sepal and petal of flowers and the response is the type of flower. The data is available from the R library datasets and can be accessed with iris once the library is loaded. It is also available at the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Iris). *The response values are only given to see how well a specific method performed and should not be used to build the model.*

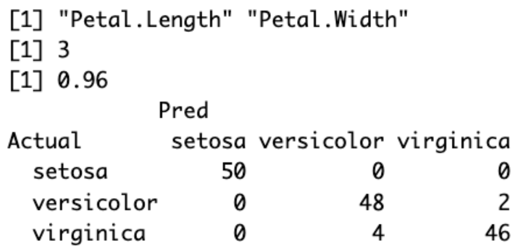
Use the R function kmeans to cluster the points as well as possible. Report the best combination of predictors, your suggested value of k, and how well your best clustering predicts flower type.

Methodology - Iris Clustering with k-means

- **Data**: 150 observations; predictors Sepal.Length, Sepal.Width, Petal.Length, Petal.Width. Species labels are used **only for evaluation**.
- **Preprocessing**: scale() to standardize features.
- **Search**: all 2-, 3-, and 4-feature combinations with k∈{2,3,4,5,6}; nstart = 50.
- **Evaluation**: map clusters → species by **majority vote**; compute **accuracy** and a **confusion matrix**.

Results - Iris Clustering with k-means

- **Best predictors: Petal.Length, Petal.Width**
- **Suggested k: 3**
- **Accuracy (after mapping): 0.96**
- **Confusion matrix (Species × Predicted):**



```
[1] "Petal.Length" "Petal.Width"
[1] 3
[1] 0.96
            Pred
Actual       setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         48         2
  virginica       0          4        46
```

Discussion - Iris Clustering with k-means

- Petal measurements cleanly separate the classes: setosa is perfectly clustered, and the remaining errors are the usual overlap.
- k=3 aligns with the three species and maximizes accuracy; larger k fragments true groups, while k=2 underfits.

Conclusion

Using stratified 5-fold CV, kNN achieved 0.7600 ± 0.0462 accuracy with k=31, outperforming the majority baseline. With a stratified 60/20/20 split, a linear SVM (C = 0.01, scaled) reached 0.8195 test accuracy and was conservative on approvals (very few false positives), reflecting a precision-over-recall trade-off. For unsupervised learning, k-means on Petal.Length and Petal.Width with k=3 aligned closely with species (0.96 accuracy), with expected versicolor–virginica overlap. Note that kNN's figure is CV accuracy while SVM's is held-out test accuracy, so the two are not directly comparable; future work could standardize evaluation, explore class-weighted SVMs, and report silhouette/ARI for clustering.

REFERENCES

https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/K-Means.