ADM CA 1 Questions: 22/4/2017 worth 20% of ADM

Instructions:

- Do all foundations exercises
- Do at least 4 other questions (your best 4 will be used to determine your score)
- Note which 3 columns you lost
- For each question, provide a screen shot of the output, and a brief discussion on the approach taken in a word document (upload this to moodle)
- Upload an executable solution for me to be able to recreate your answers (upload this to moodle)

Foundations: (worth up to 4 points)
F1: Sensibly encode the categorical attributes, and check for missing values and outliers.
F2: Inspect the dependent variable and comment on class balance
F3: Make train and test datasets (think about your answer to F2)
F4: Formulate some assumption of your data and make a 2 benchmark predictive models to test models against (think about how to use this in the CA) – e.g. no one leaves, only (wo)men leave, etc.

Basic Questions: (worth 0.75 points each)
B1: Identify which numerical attributes are correlated with each other
B2: Make 2 visualisations of the dependent variable; 1 against a categorical and 1 against a numerical attribute
B3: Make 2 gender-based observations of the dataset (if you lost the gender column use overtime, marital status, or business travel)
B4: Is there a relationship between age and hour/day/month rate (pick just one)?
B5: Pick one of the Likert scales and interpret it against the dependent variable

Intermediate Questions (worth 1.5 points each)
I1: Evaluate the performance of a C5.0 tree.
I2: Evaluate the performance of a kNN.
I3: Evaluate the performance of a logistic regression.
I4: Run k-means, find a good value of k, and interpret the results
I5: Demonstrate overfitting a model of your choice
I6: Plot the most important attributes
I7: You suddenly recover the lost 3 columns, redo Intermediate 1, 2, 3, or 4 and discuss any observed differences

Advanced Questions (worth 3 points each)
A1: Do PCA on the dataset, select the first 2 components, cluster them and visualise the output
A2: Build a SVM that significantly outperforms your F4 benchmarks – define significantly.
A3: Which is better, Naïve Bayes, a Random Forest, or a CI Forest? **Don't use only the default settings!**
A4: Demonstrate if an ANN can outperform a logistic regression on this dataset. (If you do this question, **don't do** Intermediate 3)
A5: Do some form of dimensionality reduction and/or feature engineering and demonstrate an improvement with/without it on Intermediate 1, 2, or 3.

Show Off Questions (worth up to 5 points each)
S1: Using the full dataset, beat 88.9% prediction accuracy; use the same model parameterisation on your CA dataset and discuss the differences in performance
S2: Demonstrate how changing your approach based on F2 and F3 affects 2 Advanced Questions, or 3 Intermediate Questions
S3: Automate the hyperparameter optimisation of a Deep Learner, ANN or SVM. Demonstrate that you have not overfitted, but that it outperforms at least 2 other models you have built. **(Be careful to not commit too much time to this question!!)** Some credit will still be awarded if you don't beat 2 models.

Data Description

The key to success in any organization is attracting and retaining top talent.
You are an HR analyst at my company, and one of my tasks is to determine which factors
keep employees at my company and which prompt others to leave. We need to know what
factors we can change to prevent the loss of good people.

You have data about past and current employees in a spreadsheet. It has various data
points on our employees, but we're' most interested in whether they're still with the
company or whether they've gone to work somewhere else. And we want to understand how
this relates to workforce attrition.

#Attributes:
Age: in years
Attrition: Y/N the dependent variable -- have they left the company?
BusinessTravel: Non-Travel; Traval_Frequently, Travel_Rarely
DailyRate: Consultancy Charge per Day
Department: Human Resources; Research & Development; Sales
DistanceFromHome: How far the employe lives from work
Education: 1 'Below College'; 2 'College'; 3 'Bachelor'; 4 'Master'; 5 'Doctor'
EducationField: Human Resources; Life Sciences; Marketing; Medical; Other; Technical Degree
EmployeeCount: No of employes in this record
EmployeeNumber: Employee ID
EnvironmentSatisfaction: 4 point Likert scale: 1 'Low'; 2 'Medium'; 3 'High'; 4 'Very High'
Gender: Male / Female
HourlyRate: Consultancy Charge per Hour
JobInvolvement: 4 point Likert scale: 1 'Low'; 2 'Medium'; 3 'High'; 4 'Very High'
JobLevel Metadata not available -- make an assumption ☺
JobRole: Healthcare Representative; Human Resources; Laboratory Technician; Manager; Manufacturing Director; Research Director; Research Scientist; Sales Executive; Sales Representative
JobSatisfaction: 4 point Likert scale: 1 'Low'; 2 'Medium'; 3 'High'; 4 'Very High'
MaritalStatus: Divorced; Married; Single
MonthlyIncome: monthly salary
MonthlyRate: Consultancy Charge per Day
NumCompaniesWorked: No. of previous employers
Over18: Y/N
OverTime: Yes/No
PercentSalaryHike: Last Year's Increment
PerformanceRating: 4 point Likert scale: 1 'Low'; 2 'Good'; 3 'Excellent'; 4 'Outstanding'
RelationshipSatisfaction: 4 point Likert scale: 1 'Low'; 2 'Medium'; 3 'High'; 4 'Very High'
StandardHours: Contract hours
StockOptionLevel: No available metadata -- make an assumption ☺
TotalWorkingYears: Career Age
TrainingTimesLastYear: No. of training courses attended last year
WorkLifeBalance: 4 Point Likert Scale: 1 'Bad'; 2 'Good'; 3 'Better'; 4 'Best'
YearsAtCompany: Time spent with company
YearsInCurrentRole: Time in current role
YearsSinceLastPromotion: No. of years since last promoted
YearsWithCurrManager: Years spent with current manager