Assignment # 2

Read the Telco case and answer the questions below. The questions are **bolded** to be more salient. Please make **your answers as specific as possible**. Also, whenever appropriate, use images, charts, and tables in your answers.

Background

Your team decided to try logistic regression and decision tree methods to predict who to target with retention incentives. The company has a dataset (churn.csv) that contains 7032 contract information including a set of variables describing the contract details as well as a Churn variable indicating whether this contract was renewed or not when it was expired.

Task 1: Compare the predictive performance of the two machine learning methods

Based on what you learnt so far, please use Python to compare decision tree and logistic regression methods in a comprehensive way. Please use 10-folds cross validation when you train the models. Use Accuracy, Precision, Recall and AUC as performance measures.

Question 1: what is the most effective decision tree you can build? To answer this question, you want to systematically experiment with different levels of model complexity. The easiest way is to change the "maximal tree depth" levels and leave other parameters unchanged. **Please report your findings and explain the rationale behind.** Once you settle with the most effective decision tree, **please characterize the high- and low-probability churning customers with the decision tree results. Is the output of data-driven analysis same or different from your intuitions?**

Question 2: Is decision tree or logistic regression better? Please first build a Logistic regression model with no regularization and compare with the best decision tree model you built from the previous question. **Please carefully examine various performance metrics and comment about the differences in performance between the two methods**. Next, use Lasso regularization. Please try to adjust the Strength parameter from Weak to Strong. As expected, the stronger the regularization strength is, the less variables are kept in the model. **Please compare the "good" features identified by decision tree and logistic regression and try to provide explanations about why they are different.**

Task 2: Optimize business returns by employing data-driven decision makings solutions.

You have studied how to evaluate model output with the considerations of economical cost of wrong predictions. We are going to further extend the problem solving following this line of analysis.

Cost Matrix		
	Predicted Churn (+)	Predicted Not Churn (-)
Actual Churn (+)	0	CV-205
Actual Not Churn (-)	205	0

Question 3: Reevaluate different models with the cost information and customer's contract value

One simple decision-making rule is if the model predicts a customer will churn, a retention offer will be sent to him/her. Now we take into account the differential economic benefits of sending retention offers to different customers. We assume that churning customers will actually renew a one-year contract with the same monthly payment as the previous contract, if they receive the retention offer. So, we define contract value (CV) of individual customer as: Monthly payment *12.

Also, let us make the following further assumptions. Offer itself costs $200 and sending offer costs $5 as the operational cost. If a customer is predicted not churn but he/she actually churn, the company loose CV-205, which could have been earned if the prediction is correct. Also, if a customer is predicted churn but he/she actually doesn't. However, he/she will take the offer anyway and therefore your company will waste $200 on him/her. Adding $5 of offer-sending cost, it will then cost $205 in total. The cost matrix is shown in the below Figure.

Please use the prediction output from decision tree and logistic regression model you built from Task 1 (with the default decision threshold) and **compare which model gives you better overall performance with the cost information.**

Next, if you are allowed to change the decision threshold, **what would be best decision threshold and minimal total cost generated from either model?**

In order to evaluate the true value of your data-driven solution in the previous question, we need to compare the returns with the two baselines strategies including doing nothing and sending offer to everyone, which is equivalent to predicting everyone as not churning, or predicting everyone as churning. **Please make some conclusions about whether data-driven solutions are better?**

Question 4: Making retention offer decisions based on calculating the expected return.

Now, we further extend our analysis by using the expected value framework (Check Expected value - Wikipedia, if necessory). That is, we only send to customers whose expected return is positive, which means that their contract value (CV) multiplied by their probabilities of churning is worth more than retention cost ($200) and offer transaction cost ($5). That is, the retention offer decisions are conditional on:

$$CV * Prob(churning) - 205 > 0$$

For the probabilities of churning for each customer, you are going to use those generated by the best decision tree and logistic regression models from the first task.

Again, you can evaluate your retention offer decisions with the same cost matrix. **Please report the evaluation results.**

Question 5: The assumption that the churning customers will not churn if they receive the retention offer might not be realistic. Please make a short proposal about how to further improve the solution.

Deliverables: A Word file that contains your answers to the above questions and a Python file that contains your codes.