BUDT733 – Data Mining

"Tourism Insurance: Predicting Days with Unhealthy Air Quality in Washington, D.C. "

May 12, 2008

Team 1
Kate Hannon
Camille Hoff
Tian Liu
Clayton Paulding
Ramin Riahikhoee

Executive Summary

***Business Problem**:* D.C. Ducks is a tour company whose revenues depend heavily on favorable weather conditions, such as air quality. On days when the air quality index (AQI) is above 50, the firm has fewer customers due to the adverse conditions. The firm currently purchases a blanket weather insurance policy for the entire year, as protection against days when the AQI is forecast to be above 50.

***Task**:* Predict dates on which the Air Quality Index is above 50, for the purpose of selectively purchasing weather insurance for these days.

***Data**:* Historical weather data from the past three years, a total of 1096 records each containing 17 variables. [Note: This task should have used historical forecast data; however the necessary data were unavailable, see Technical Report for further details.] Variables described typical data about the weather such as temperature, precipitation, visibility and wind speed. AQI levels for each day were collected for use as the response variable.

***Model Creation and Selection**:* A classification tree was selected from a variety of logistic regression, discriminant analysis and classification tree models. The classification tree was selected based on its error rate (the lowest we found), parsimony and ease of explanation. The final model gave an error rate of approximately 22% on the validation data. The misclassification rate for the "expensive" errors i.e. misclassifying a high AQI day as a normal business day was just above 14%.

***Analysis of Results and Recommendations**:* We conducted a cost benefit analysis to accurately measure the benefits of purchasing insurance using our classification tree instead of buying blanket insurance. We believe that by using the classification tree to selectively purchase insurance, DC Ducks can save almost $40,000 annually!

Technical Summary

***Business Problem*:** D.C. Ducks provides guided tours on land and water in Washington, D.C. throughout the year. On days with an air quality index (AQI) above 50, few tourists are interested in taking a D.C. Ducks tour, and the company's daily revenues fall to almost zero. As a remedy, D.C. Ducks purchases weather insurance for every day of the year at a cost of $70 per day (**Exhibit A**). On days with an AQI above 50, the insurance company pays D.C. Ducks its average daily revenue of $13,720. This insurance policy protects D.C. Ducks from some unexpected revenue losses, but at a substantial cost. D.C. Ducks management would like to predict days with AQI above 50 in advance, so it can purchase weather insurance only during the limited time periods that it makes economic sense. The cost of insurance for selected dates is $100 per day.

***Note on Data*:** We obtained actual historical weather and AQI data for the Washington, D.C. area. Since D.C. Ducks wanted a model to apply to weather *forecasts* to predict future days with AQI above 50, these data were not appropriate for this predictive task. Ultimately we were forced to base our model on *actual* weather observations, instead of *forecast* data since finding past *forecasts* proved impossible. In other words, it would have been more accurate to base our model on historical weather *forecasts* as opposed to historical *actual* data. Ultimately we decided to divide our actual weather data into two categories: training (April 1, 2005-March 31, 2007) and validation (April 1, 2007-March 31, 2008) data. Then, we conducted our modeling as if all the data were forecasts and not records of actual past weather conditions. While this was not ideal, we felt that it was the best solution available for our client.

***Data Processing and Exploration*:** After partitioning the data, we visually explored it in Spotfire to see which variables displayed notable patterns and relationships. We quickly concluded that temperature variables remain the dominant differentiators of days with high and low AQI. One of the more interesting patterns was that days with AQI above 50 occurred more frequently when the day's actual high temperature was significantly higher than the day's historic average temperature (**Exhibit B**). Further exploration showed that a few other variables also contribute to data separation in meaningful ways: low precipitation, lower dew point averages, low average or

maximum wind speeds and higher visibility levels all indicate days with high AQI levels. However, these variables' separation power is much more ambiguous than temperature. In addition, we initially thought of the data as time series data but soon realized that we could drop time information from the data analysis. Days with an AQI above 50 were more frequent in the summer, but they still occurred occasionally in winter. We concluded that date was heavily correlated with temperature, which seemed to be the dominant factor in determining days with AQI above 50. We realized that whether the day was in summer or winter was irrelevant. Indeed, the model would be less useful if it attempted to capture the time series nature of the raw data, because it would need to incorporate strong seasonality.

***Model Creation and Selection*:** After exploring the data, we developed discriminant analysis, logistic regression, and classification tree models. We quickly rejected the discriminant analysis models, because their error rates were well above 60%. We next turned our attention to the logistic regression models. Our first model incorporated the seven variables that our data exploration had shown to be significant. We then built a second all-inclusive model with 12 variables and compared the two. We finally settled on a logistic regression model with five variables (average temperature, average dew point, average visibility, average wind speed, and precipitation), all of which had very low p-values (**Exhibit C**). The error rates for these three logistic regressions models were similar (around 25%) but we felt that in the interest of parsimony, the model with the fewest variables was best. The classification tree had a slightly lower error rate (at 23.22%) than both logistic regressions and was easier to explain (**Exhibit D**). This offered a big advantage since D.C. Ducks employees are unfamiliar with data mining techniques. The tree used slightly different variables than the ones used in the regressions e.g. High instead of Average Temperature. The tree became our preferred model for the task.

***Results and Recommendations*:** We used past financial data to quantify the cost of misclassification, and then used the error rates from our classification tree to obtain an estimate of the potential cost savings from using this model. If D.C. Ducks can purchase insurance for only those days during the year that it predicts will have an AQI over 50, they can save over $38,000 per year (**Exhibit E**).

Exhibit A: Weather Insurance Breakdown

Daily expected revenue	$13,720
Cost of blanket insurance, per day	$70
Revenue from insurance collection, per day	$13,720
Number of Days in 3 Year Data Period (1 leap year)	1,096
Historical Number of High AQI Days in 3 years	379
Cost of blanket insurance (insure every day for 3 years)	$76,720
Expected 3-yr revenue, blanket insurance	$15,037,120
Expected 3-yr profit, blanket insurance	$14,960,400
Expected 3-yr profit, no insurance	$10,656,324

Exhibit B: Scatter plot of Average versus High Temperature and AQI Above 50

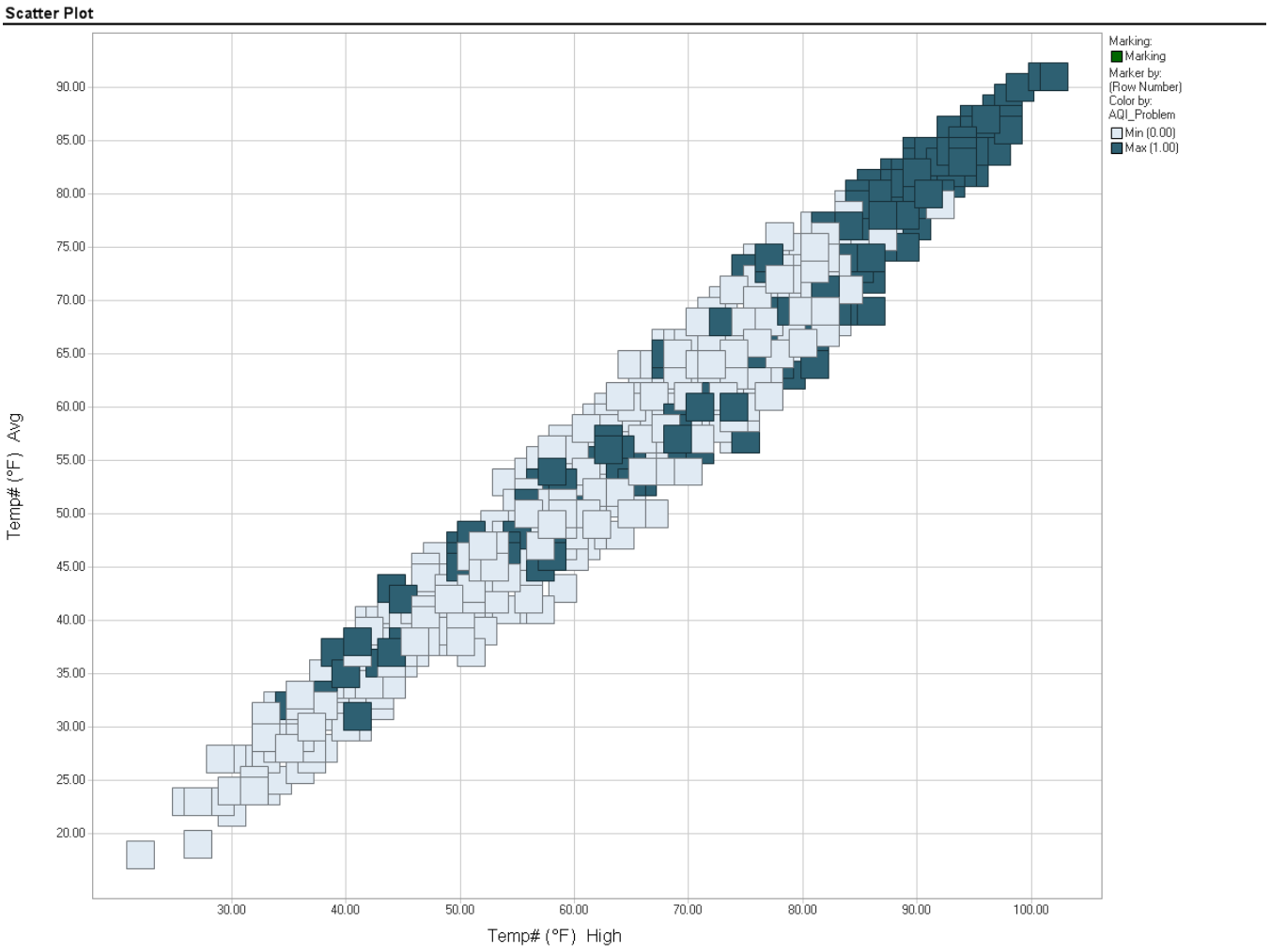

Exhibit C: Logistic Regression Model -Validation Data Summary Report and Variables

Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	89	55
0	37	185

Error Report			
Class	# Cases	# Errors	% Error
1	144	55	38.19
0	222	37	16.67
Overall	366	92	25.14

Input Variables	Coefficient
Constant Term	-1.2581526
Temp. (°F) Avg	0.22644049
Dew Point Avg	-0.11404138
Visibility Avg	-0.72533727
Wind Avg	-0.21747842
Precipitation (inches)	-1.96287334

Exhibit D: Predicted versus Actual Summary Report for the Classification Tree

Validation Data scoring - Summary Report (Using Best Pruned Tree)

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	91	53
0	32	190

Error Report			
Class	# Cases	# Errors	% Error
1	144	53	36.81
0	222	32	14.41
Overall	366	85	23.22

Exhibit E: Potential Cost Savings from Buying Insurance for only Predicted AQI Days

	# of Days (in 3 yrs)	Total Value (3 yrs)
Normal AQI days correctly predicted	569	$7,806,155
High AQI days correctly predicted	273	$3,719,662
High AQI days incorrectly predicted as normal	159	($2,177,506)
Normal days incorrectly predicted as High AQI	96	$1,308,013
Cost of blanket insurance, per day	$70	
Cost of selective insurance, per day	$100	
Cost of blanket insurance (insure every day, 3 years)		$76,720
Cost to insure 379 selected days over 3 years		$37,900
Expected 3-yr Profit, blanket insurance		$14,960,400
Expected 3-yr Profit, selected days insurance		$14,999,220
Benefit of selected days insurance over blanket insurance:		**$38,820**