Analysis of Classifier Designs for Flight Cancellation Prediction UCLA - Statistics 101C - Fall 2020 Group 1H

Emily Allendorf
UCLA
emilyjallen@g.ucla.edu

Nurrachman Liu
UCLA
rachliu@gucla.edu

Avyay Kuchibutla
UCLA
avyayk@ucla.edu

***Abstract*— We explore a range of classifier designs for the prediction of flight cancellation. We found the best classifier design was PCA+SVM (99.869%/99.943% public/private), followed closely by random forest (99.83%) and SVM (99.65%). The differences between all of these were small (<1%) but the strongest model also included merged precipitation data from the Infoplease "Climate of 100 Selected U.S. Cities" data set. Additionally,the difference between top classifiers and other linear classifiers were quite large (>20%) suggesting that the binary response classes are separated by a highly non-linear decision boundary.**

Our final Kaggle rank for our best classifier, PCA+SVM, was rank 1 (3-way tie amongst 3 groups) on the private leaderboard (99.943%), and rank 5 on the Kaggle public leaderboard (99.869%). This suggests our final classifier, PCA+SVM, was quite robust and not as overfitted as other groups, since it performed even better on the privately withheld portion of the test-set.

I. INTRODUCTION

Flights are often prone to cancellations, incurring lost resources and economic penalties. The ability to accurately forecast their cancellation, then, is important for mitigating their economic impact.

We investigate a range of standard classifier designs and their ability to predict flight cancellation. We present designs for several classifiers, and analyze their efficacy. We then discuss the design process and trade-offs between classifiers for this dataset.

II. OVERVIEW

The paper is organized as follows:

1) Methods: parameters and scope of the data and analysis.
2) Pre-modeling data analysis: building intuition of the dataset using pairwise scatters.
3) Understanding our predictors: use of some results from modeling to explain our predictors.
4) Phase 1 Models: brief analysis and results for each fitted classifier.
5) Phase 2 Models: analysis of most successful models following the addition of new weather predictors
6) Discussion: classifier comparisons; threshold selection.

III. METHODS

A. Kaggle

This project was performed as a class contest on Kaggle. There is a provided training set (69,000 obs) of 50 predictors. The test-set (30,000 obs), minus the labels, is provided. The goal is to train a model to predict the test-set labels.

B. Dataset

The data set used was Flight predictions. There were 50 predictor columns. They can be separated into the following categories: *airline metrics, geographical metrics, and time metrics*.

Our second round of models also include two additional predictors which fall into the category of *weather* coming from an Infoplease climate data set [1].

C. Models

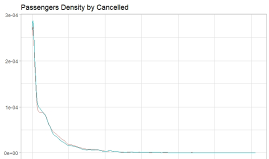
We analyzed the following models:

1) Logistic Regression
2) Discriminant Analysis (LDA and QDA)
3) K Nearest-Neighbors
4) PCA + Logistic Regression
5) Decision-Tree
6) Random Forest (Bagging)
7) Boosted Trees
8) SVM
9) PCA + SVM

IV. PRE-MODELING DATA ANALYSIS

A. Data Cleaning

Histograms of certain predictors (e.g. 'passengers', 'Median Income') suggest they were joined on the airport field, due to their low number of unique values and missing value ratios of as high as 97%. This made us doubt their usefulness, as unique airports (200+) number much fewer than total observations (69,000+), and this was confirmed by low separability in density plots.



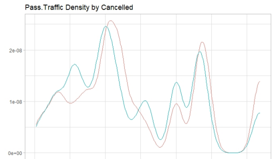


Fig. 1: **Low separability parameters.** include 'Passengers' and 'Pass.Traffic' (shown), race and ethnicity parameters, and delay parameters.

From scatter plots, geographic latitude and longitude provide good separability for cancellation. State predictors had too many levels (51) but contained crucial information. After re-leveling into four-tiered 'region' predictors (West, East, South, and Midwest), O.Region and D.Region were very useful, especially for explainability.

B. Single Dimension Exploration

We use pairwise plots to understand our data; these are separated into 3 plots below.

In each plot, the response is shown as the last column and last row. The diagonals show their distributions. Each predictor's relationship to the response is along the very last row of each graph.

Figure 2 shows geographic predictors. **We found these to provide the most separability, and therefore, the most useful.** For example, the 'East' region is a good predictor of cancellation. Longitude and latitude provide moderate separability.

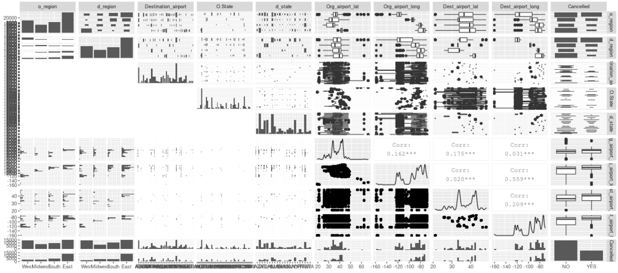


Fig. 2: **Geographic metrics scatters.** Predictors from left to right: passengers, seats, flights, distance, origin_population, destination_population, airline, cancelled (response).

'Distance' has low correlation to cancellation. This makes sense given that distance cannot capture geographic difference in terms of region.

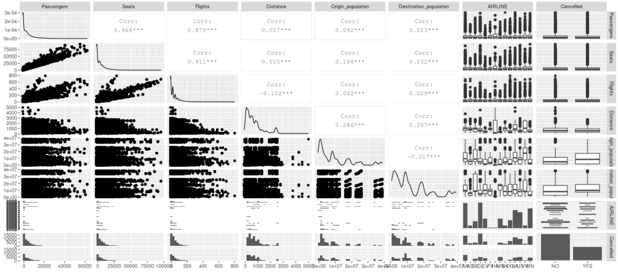


Fig. 3: **Airline metrics scatters.** Predictors from left to right: o_region, d_region, destination_airport, o_state, d_state, orig_airport_lat, orig_airport_long, dest_airport_lat, dest_airport_long, cancelled.

Figure 3 shows the predictors for airline metrics, such as the passengers per airline. We observed that none of predictors for airline metrics were useful, and therefore did not use these predictors.

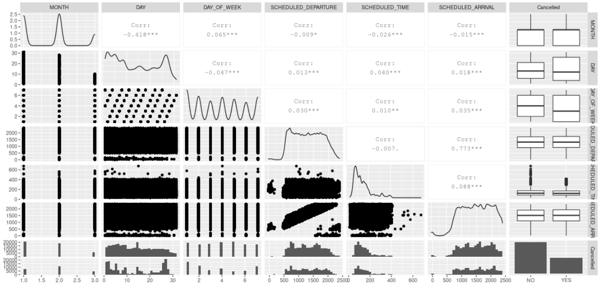


Fig. 4: **Time metrics scatters.** Predictors from left to right: month, day, day_of_week, scheduled_departure, scheduled_time, scheduled_arrival.

Figure 4 shows that 'Day' provides good separability. Holidays around 20-30 and 1-5 provide particularly good signal. 'Day of the week' and 'Month' provide moderate separability. Therefore, the time-metric predictors were the second-most important set of predictors.

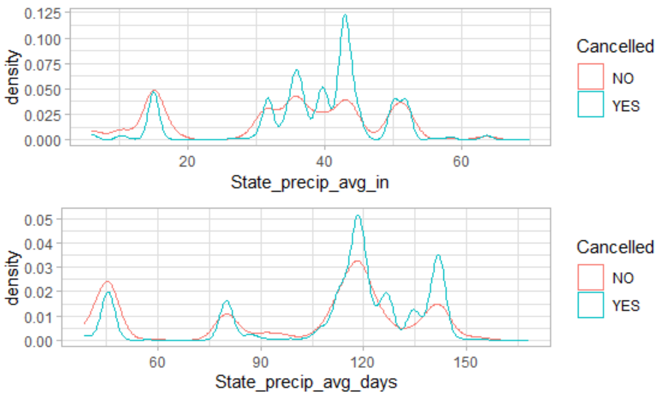


Fig. 5: **Weather Predictor Density Plots.** Predictors from left to right: State_precip_avg_in and State_precip_avg_days.

Figure 5 shows the densities of the last two predictors we considered- and the ones we added to our second round models: annual precipitation. The separability seen was the strongest among all the city-level weather predictors in the City Weather data set from infoplease.com.

C. Predictor Rankings based on Single-Dimension Exploration

Based on the above, we determined that the best predictors were, in estimated order of importance:

1) o_region
2) d_region
3) o_lat, d_lat
4) d_long
5) o_long
6) State_precip_avg_in
7) State_precip_avg_days
8) day
9) day_of_week
10) month
11) distance

These preliminary predictor rankings turned out to be quite accurate. However, depending on the model, precipitation predictors outperformed latitude and longitude.

V. Understanding our predictors

We borrow some of our PCA graphs to help understand our original predictors. We found that the time-metric predictors vs geographic-metric predictors are entirely aligned on PC2 and PC1, respectively; implying that there isn't much to be gained from doing PCA (see biplot, Fig. 6).

More importantly, it implies that our original predictors are already able to explain the maximum variation very efficiently (ie, with low VIF).

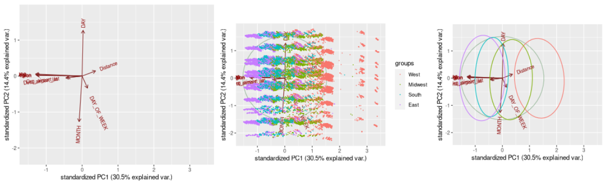


Fig. 6: PC1 vs PC2: biplot (left), scatter (middle), ellipsed (right)

Fig. 7 shows the test-set data rotated onto the PC axes and overlaid on top of the training data. The same regions in test and training closely overlap. This provides explanation for the very tight correlation between training and test accuracy.

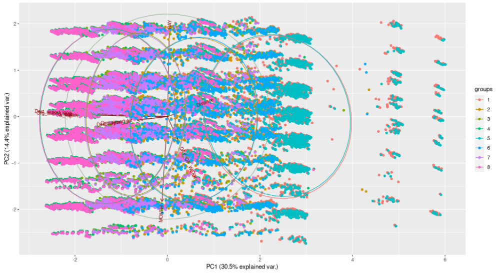


Fig. 7: Test-set predictors rotated onto PC1 and PC2.

VI. Round I Models

A. Formulae

We used the following formulae in our various models; they are numbered consistently as follows:

1) Cancelled ∼ o_region + d_region + Distance + Org_airport_lat + Dest_airport_lat + Dest_airport_long + Org_airport_long + DAY + DAY_OF_WEEK + MONTH
2) Cancelled ∼ o_region + d_region + Distance + Dest_airport_lat + Dest_airport_long + Org_airport_long + DAY + DAY_OF_WEEK + MONTH + SCHEDULED_TIME + SCHEDULED_DEPARTURE
3) Cancelled ∼ d_region + Distance + Dest_airport_lat + Dest_airport_long + Org_airport_long + DAY + DAY_OF_WEEK + MONTH + SCHEDULED_TIME + SCHEDULED_DEPARTURE

B. Logistic Regression

The logistic regression performed poorly. All of the fitted predictors were fitted with high statistical significance, confirming their choice, yet this classifier only achieved an accuracy of 72%. x The ROC and PR curves for the logistic regression classifier:

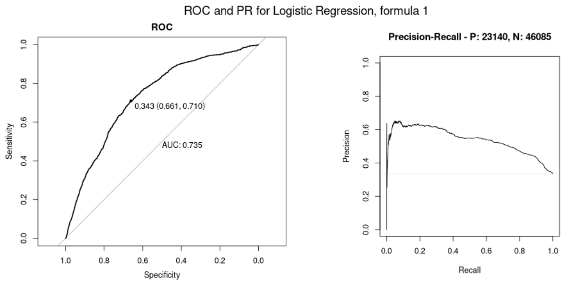


Fig. 8: ROC (left) and PR (right) for the logistic regression model, formula 1

C. Discriminant Analysis (LDA and QDA)

Neither LDA nor QDA perform well. Their accuracies were 70% and 75%, respectively. The ROC AUCs between LDA (0.722) and QDA (0.761) are comparable (Fig. 9).

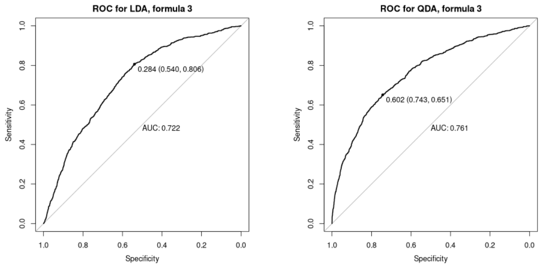


Fig. 9: ROC curves for Gaussian Discriminant Analysis, LDA (left), QDA (right).

The scaled group-means for LDA confirm that there is too much overlap and spread to allow for strong separation.

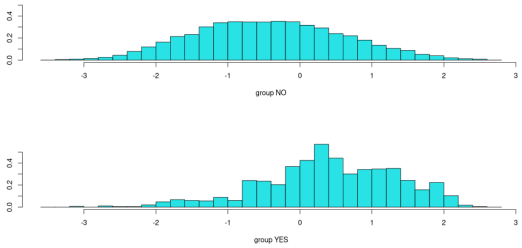


Fig. 10: Group-means for LDA.

D. k-Nearest Neighbors

The kNN accuracy is $>90\%$ for k < 10, leveling off at 80% for extreme k values of >100.

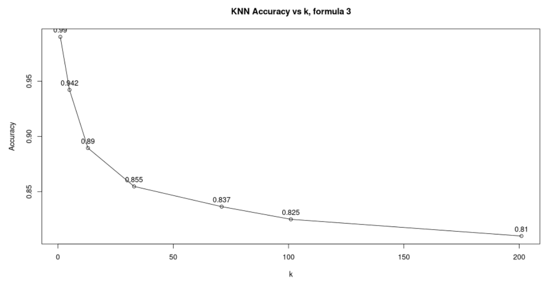


Fig. 11: Accuracy vs k for k-Nearest Neighbors, formula 4.

Low-k KNN is typically at risk of overfitting; however, in this case, it consistently scored the best on the public test-set. One possible explanation is that, with ten predictors, the distances are large enough to avoid overfitting, even at k=1. Another possible explanation is that there is enough data such that the decision boundary is relatively continuous and free from noise caused by gaps, which would present severe noise for k=1.

E. PCA

For PCA, we scaled the predictors to ensure that high-variance variables do not get over-emphasized [2].

For visual analysis of the PC components, we cluster (color) by 'region'. See Fig. 6 (shown earlier) for PC1 vs PC2.

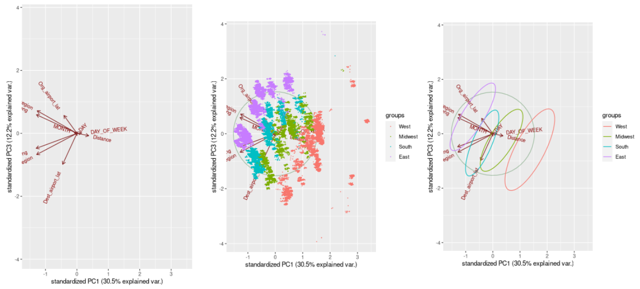


Fig. 12: PC1 vs PC3: biplot (left), scatter (middle), ellipsed (right)

PC1 vs PC2 was not able to separate the regions clearly (the ellipses overlap greatly). PC3 is able to give complete separation between the ellipses. Both PC3 and PC2 explain 12.2% of the variation.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	1.7461	1.1983	1.1032	1.0463	0.9893	0.9774	0.7600	0.7239	0.3124	0.2645
Proportion of Variance	0.3049	0.1436	0.1217	0.1095	0.0979	0.0955	0.0578	0.0524	0.0098	0.0070
Cumulative Proportion	0.3049	0.4485	0.5702	0.6797	0.7776	0.8731	0.9308	0.9832	0.9930	1.0000

TABLE I: PCA cumulative variation explanations, formula 1

Table I shows that a model shrinkage of 3 principal components (10 to 7) retains 93% variation explained.

a) PCA+Logistic Regression: There was no improvement PCA+logistic regression. This makes sense, given that our major predictors align closely on the PC1 or PC2 axes.

b) PCA+Tree: There was no improvement The PCA+Decision-Tree model only reduces the tree-depth by 1.

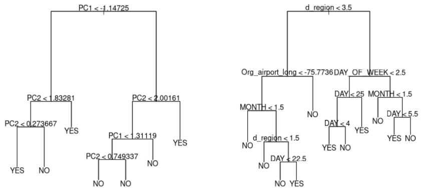


Fig. 13: Decision-Tree for PCA+Decision-Tree (left) vs original Decision-Tree (right).

F. Decision-Tree

The following is our fitted decision-tree model.

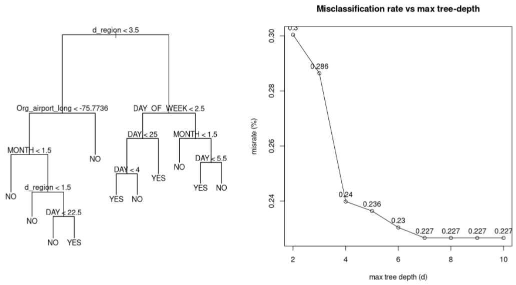


Fig. 14: Fitted decision tree (left); Accuracy vs depth (right).

Tree accuracy increases with depth, leveling off to accuracies of 77% beginning at depth 7. Its ROC AUC is 0.78 with the opt-threshold at 0.308.

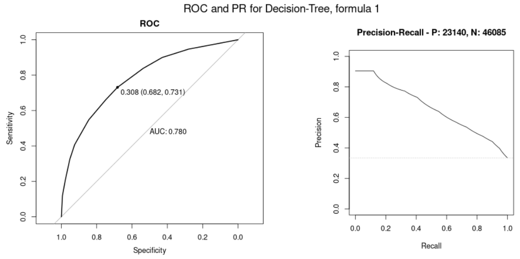


Fig. 15: ROC and Precision-Recall for Decision-Tree, formula 1

Thresh = 0.5 (acc = 0.752%)		
pred/actual	NO	YES
NO	42624	13720
YES	3461	9420

Thresh = 0.308 (acc = 0.699%)		
pred/actual	NO	YES
NO	31447	6222
YES	14638	16918

TABLE II: Tree model confusion matrices for 0.5 threshold and best (0.308) threshold, formula 1.

G. Random-Forest (Bagging with p=5)

The random-forest model gave us the highest accuracy. Its variable importance plot ranks our chosen predictors (Fig. 16).

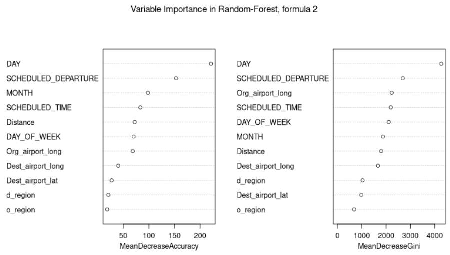


Fig. 16: Random-forest variable importance, formula 2

Formula 2 improved the accuracy from 99.03% to 99.65% (99.802% on public). The improvement supported the strength of adding SCHEDULED_TIME and SCHEDULED_DEPARTURE to our model.

Since the algorithm outputs hard classifications rather than scores, the equivalent parameter to threshold was the votes at the leaves. However, we found empirically that the votes were very decisive and so provided no tunability.

H. Boosted Trees

Boosted trees also achieved very high accuracy, as high as kNN (98+%) and almost as high as random-forest (99%). However, the main caveat is that the tree depth must be very high (> 10 for 95% acc, and > 18 for 98% acc).

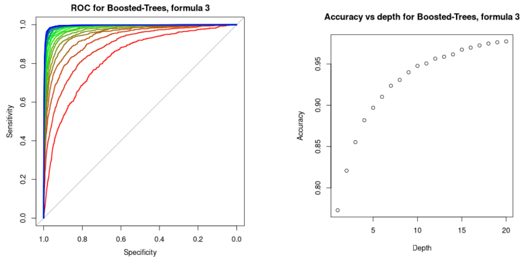


Fig. 17: Boosted-Trees ROC with varying depth (left), Accuracy vs depth (right), formula 3. ROC curve: red is min depth (1) and blue is max depth (20).

I. SVMs

We fit the four main kernels of SVM (radial, linear, polynomial, and sigmoid), with varying hyperparameters for each (Fig. 18).

Radial performed far better than all other SVM kernels (99.2%), at around the level of the random-forest (99.6%). Linear performed very poorly (66%) at no-skill. Polynomial peaked at degree 5 (84%).

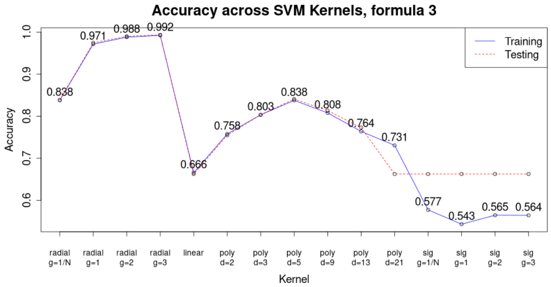


Fig. 18: Accuracy for different kernels, SVM, formula 3.

In polynomial space, the best fit was at degree 5. This may help explain why the linear classifiers, and QDA, perform poorly.

SVM with radial kernel is able to achieve near optimal classifier performance (Fig. 19).

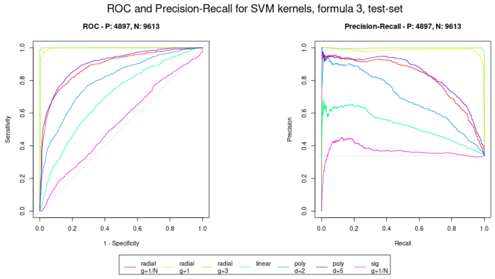


Fig. 19: ROC and Precision-Recall for SVM with varying kernels, formula 3. Test-set. Sigmoid (magenta); linear (turqouise); polynomial (blue to purple); radial (red, yellow, green).

VII. Round II Models

Round II built off the most successful models -namely SVM and Random Forest. From here on, we will proceed with the following formula:

$$\text{Cancelled} \sim \text{d\_region} + \text{o\_region} + \text{Distance} + \\ \text{Dest\_airport\_lat} + \text{Dest\_airport\_long} + \text{Org\_airport\_lat} + \\ \text{Org\_airport\_long} + \text{DAY} + \text{DAY\_OF\_WEEK} + \text{MONTH} + \\ \text{SCHEDULED\_TIME} + \text{SCHEDULED\_DEPARTURE} + \\ \text{State\_precip\_avg\_in} + \text{State\_precip\_avg\_days}$$

This formula amounts to a full model of all the predictors considered to this point as well as the two precipitation predictors.

A. Linear Classifiers

An initial comparison with the linear classifiers (logistic regression and PCR) confirmed they were still far poorer than the nonlinear and non-parametric classifiers, but did show some improvement. Logistic regression and PCR both improved accuracy and sensitivity but worsened Specificity. ROC AUC was not greatly improved for either model (0.735 vs 0.738 in both cases). Additionally, PCR still separated the numerical predictors by region (seen in Figure 20), and showed promise for combinations with other nonlinear classifiers.

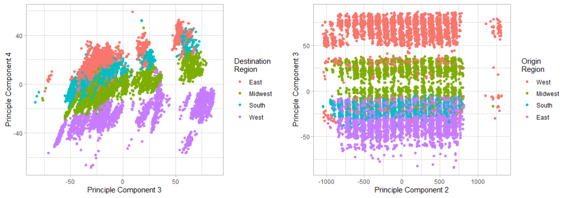


Fig. 20: Scatter plots of principle components including the weather data and colored by region variables (destination on the left and origin on the right)

Classifier	ROC AUC	Thresh=0.5 ACC / PPV TPR / FPR	Thresh=opt ACC / PPV TPR / FPR	opt thresh
Logistic Regression	0.735	70.5% / 60% 36% / 12%	68% / 51% 71% / 34%	0.343
LDA	0.722	71% / 60% 37% / 12%	62.5% / 46% 80% / 46%	0.284
QDA	0.761	68% / 52% 73% / 35%	71% / 56% 67% / 26%	0.602
PCA + Log. Regr	0.718	70.5% / 61% 34% / 11%	65% / 49% 73% / 39%	0.335
PCA + D-Tree	0.765	75% / 66% 53% / 14%	75% / 66% 53% / 14%	0.456
D-Tree	0.780	75% / 73% 41% / 7.5%	70% / 83% 73% / 32%	0.308
KNN, k = 1	n/a	99% / 97.4% 99.7% / 1.4%	n/a	n/a
KNN, k = 11	n/a	90% / 79% 95.4% / 13%	n/a	n/a
KNN, k = 101	n/a	83% / 75% 72% / 12%	n/a	n/a
Random-Forest d=1	n/a	93.6% / 94.1% 86.4% / 2.7%	n/a	n/a
Random-Forest d=5	n/a	99.9% / 99.7% 99.93% / 0.14%	n/a	n/a
Random-Forest d=10	n/a	99.8% / 99.4% 99.94% / 0.28%	n/a	n/a
Boosted-Trees d=1, n=100	0.801	75% / 74% 41% / 7.2%	72% / 55% 79% / 32%	0.325
Boosted-Trees d=5, n=100	0.884	82% / 80% 61% / 7.4%	81% / 69% 79% / 18%	0.359
Boosted-Trees d=20, n=100	0.964	90.5% / 88.9% 81.5% / 5.1%	90.2% / 81.5% 91.5% / 10.4%	0.348
Boosted-Trees d=20, n=500	0.996	97.8% / 96% 97.4% / 2.0%	97.7% / 95.3% 98% / 2.5%	0.454
SVM - sigmoid g=1	0.508	66.3% / 0% 0% / 0%	38.5% / 33.1% 80.0% / 82.5%	0.334
SVM - linear	0.698	66.3% / 0% 0% / 0%	46.2 %/ 37.9% 92.6% / 77.4%	0.323
SVM - poly d=2	0.801	75.5% / 68.2% 51.1% / 12.1%	72.8% / 57.2% 76.9% / 29.3%	0.310
SVM - poly d=5	0.904	84.1% / 82.7% 66.9% / 7.1%	82.1% / 69.2% 84.7% / 19.2%	0.231
SVM - poly d=9	0.915	81.3% / 91% 49.6% / 2.5%	81.7% / 67.7% 87.5% / 21.3%	0.261
SVM - radial g=1/N	0.887	83.9% / 79% 71.1% / 9.7%	83.5% / 76.2% 74.4% / 11.9%	0.425
SVM - radial g=1	0.995	97.5% / 95.7% 96.9% / 2.2%	97.4% / 95.1% 97.3% / 2.5%	0.422
SVM - radial g=3	1.000	99.3% / 98.6% 99.5% / 0.74%	99.4% / 99.0% 99.3% / 0.52%	0.766

TABLE III: Round I Classifier results.

B. Random Forest

We proceeded with Random Forest using the strongest classifier from Round I with a depth of 5.

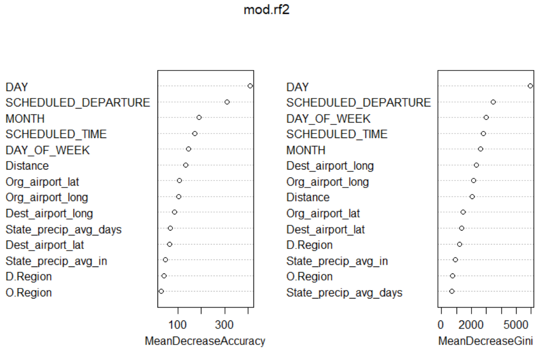


Fig. 21

As Figure 21 suggests, the precipitation predictors complete with D.Region and O.Region for strength.

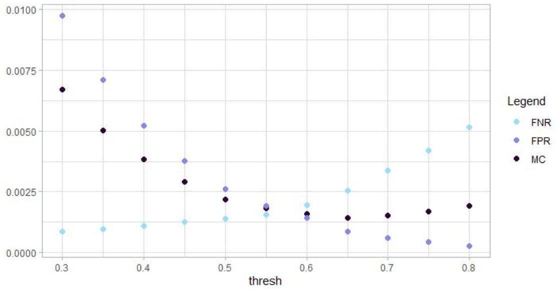


Fig. 22

This round we also ran the model to allow for changing the classification threshold so as to better optimize along different error rates. Optimizing for a minimized MC and FPR suggests a threshold anywhere from 0.6 to 0.65 would perform better than the Round I model, but testing and submitting to Kaggle with the threshold 0.65 did not perform better than a comparable round I model (0.99523 and 0.99802 respectively).

C. PCA + SVM

Inspired by the PCA improvements, we skipped straight to joining PCA with the strongest SVM model in round I, which used a radial kernel and a gamma of 3. To great success, the model outperforms previous models by one to two powers of ten, regardless of threshold.

Again, we considered classification thresholds when training and testing our model. Several thresholds were submitted to Kaggle, and the highest performing used the automatic threshold chosen by the SVM function (0.99869 public score). The nearest threshold to this was .5 which scored 0.99865 publicly.

	Round 1				Round 2			
	Acc (%)	Spec (%)	Sens (%)	Opt Tresh	Acc (%)	Spec (%)	Sens (%)	Opt Tresh
Logistic	68	66	71	34	70.3	56.8	75.3	45
PCR	65	61	73	34	70.3	56.8	75.3	45
Random Forest (d=5)	99.7	99.8	99.9	-	99.85	99.9	99.7	65
PCA + SVM (radial, g=3)	-	-	-	-	99.975	99.967	99.991	50

TABLE IV: Summary of Round II Models

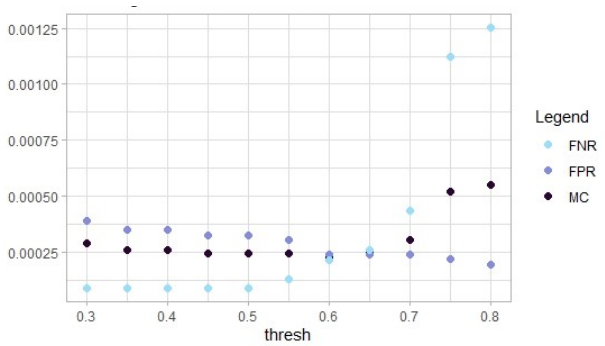


Fig. 23: Type I (lavender), Type II (light blue), and misclassification (dark purple) training errors for each threshold from the PCA+SVM Model.

Looking at Figure 23, a threshold of .5 maximizes sensitivity over specificity. The point at which all points are overlapping is at a threshold of .6 and it is also at this point where accuracy is maximized. Therefore, choosing a model based on trying to minimize all three values at once was not a successful strategy for this model.

We summarize the results of our Round II models in Table IV

VIII. Discussion

A. Classifier Comparison

Following our two rounds of modelling (summarized in Tables III and IV we have a series of observations about types of classifiers and the metrics by which we most successfully compared them:

a) Among the classifiers which performed poorly:

1) Simple decision trees performed poorly, despite Boosted-Trees and Random-Forest doing so well. However, since PCA+D-Tree does no better, the problem may not be with the decision boundaries being axis-aligned.
2) The dataset had a N:Y class balance of 2:1, and so the no-skill floor was at 66% accuracy. In this context, the linear classifiers perform quite poorly.

b) Among the classifiers which performed well:

1) The success of PCA combined with SVM suggests a linear relationship between the predictors and a nonlinear decision boundary. A compounded strategy like this is more flexibly and shows potential to be applied to different data sets with similar success.
2) While SVM and Random-Forests are powerful, they give low explainability compared to other models like kNN
3) Our regional variables ('o.region' and 'd.region') were ranked highly by Random Forest, which tells us our powerful model was also fairly simple containing just two categorical variables which four levels each.
4) As SVM can take much longer to train, linear SVM kernel should be tried only if a linear classifier such as logistic regression can perform well.
5) PCA and its associated visualizations aid in gaining intuition over the predictors. For example, the biplot provides quick feedback on how the predictors interrelate. Furthermore, it's success in separating based on region inspired its use with SVM to great success.

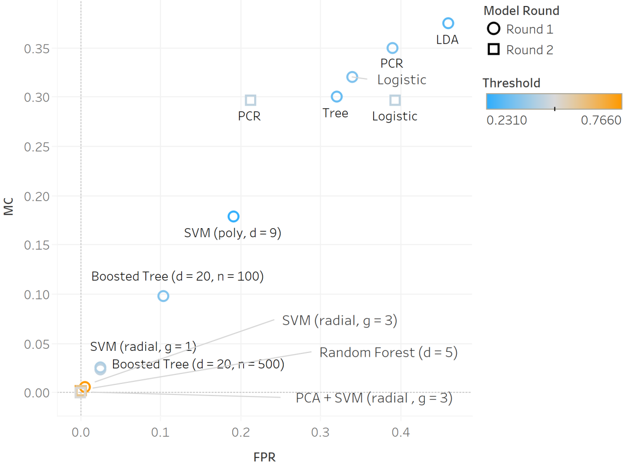


Fig. 24: Misclassification rate by false positive for all models colored by optimal threshold and shaped according to the round during which they were created and tested.

c) Regarding thresholds and choice of error:

1) Optimum threshold depends on the model as well as the situation-specific weighting of sensitivity versus specificity. For example, the threshold .5 for the PCA+SVM model performed best on Kaggle while minimizing false negatives and maximizing sensitivity. However, airlines might be seeking to minimize false positives, making a threshold of .65 more optimal.
2) optimum thresholds for our most successful models were consistent with the prior Y:N balance around .5 to .7 (see Figure 24).
3) Optimizing for one type of error (misclassification, Type I, Type II) also depended on the model.
4) opt-threshold for maximizing TPR and FPR tends to the inverse of the Y:N balance. This is explained in the following subsection.
5) The amount of imbalance between FPR and FNR cor-

relates with the distance between the chosen threshold and opt-threshold.

B. Threshold Selection

We are interested in the impact of threshold selection on accuracy. Accuracy is defined as:

$$\begin{aligned}\text{acc} &= \frac{\text{misclassifications}}{\text{total cases}} \\ &= \frac{TP+TN}{\sum \text{positive cases} + \sum \text{negative cases}} \\ &= \frac{TPR \cdot \sum \text{positive cases} + (1-FPR) \cdot \sum \text{negative cases}}{\sum \text{positive cases} + \sum \text{negative cases}}\end{aligned} \tag{1}$$

So, in balanced datasets, where the two classes are equal in size, tuning threshold and thus TPR and FPR gives equal trade-off in accuracy.

$$\begin{aligned}\text{acc} &= \frac{TPR + (1-FPR)}{2} \\ &= \frac{1}{2} \cdot TPR + \frac{1}{2} \cdot (1-FPR)\end{aligned} \tag{2}$$

However, our dataset is unbalanced with a ratio of 2:1 in favor of class 'No':

	Cancelled
NO	46085
YES	23140

TABLE V: Class balance for response 'Cancelled'

This means our accuracy is weighted approximately as:

$$\text{acc} = \frac{1}{3} \cdot TPR + \frac{2}{3} \cdot (1-FPR) \tag{3}$$

This explains why accuracy was lost when moving from a threshold of 0.5 to optimal thresholds that are < 0.5; increasing TPR is not as effective as simply decreasing FPR, when N:P > 0.5.

Eqn. 3 also shows what happens when threshold is set to the two extremes: one term dominates, converging towards an accuracy of 1/3 or 2/3 (e.g., Fig. 25).

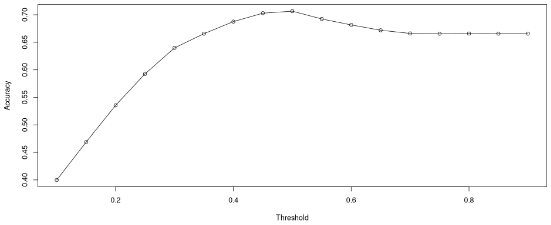


Fig. 25: Accuracy vs Threshold Trade-off for Logistic Regression, formula 3.

If N:Y is very skewed, total accuracy would not suffice, due to a high no-skill floor. Precision-Recall curves would be thus be more apt than ROC curves.

IX. Conclusions

We explored a range of linear and nonlinear, parametric and nonparametric classifiers for predicting flight cancellations from the given data set. In Round I, we found the best performing classifiers to be random-forest, boosted-trees, radial SVM, and KNN. In Round II, we added to the given dataset by merging in city-level weather data from the Infoplease.com "Climate of 100 Selected U.S. Cities' data set. Building off of Round I models, we found Random Forest to be more successful, but our most successful model was integrating PCA and radial SVM. In both phases, our most successful classifiers were those which are able to model non-linear decision boundaries, and they severely outclassed the linear classifiers such as logistic regression and LDA. We used PCA analysis to help understand the relationships of our original predictors, and analyzed the impact of threshold selection on accuracy, specificity, and sensitivity. We were also able to get a sense of the classifiers against each other, through comparing ROC curves, as well as gauge their own performances, through PR curves.

ACKNOWLEDGMENTS

Thank you to Professor A. Almohalwas.

References

[1] *Climate of 100 Selected U.S. Cities*. en. URL: `https://www.infoplease.com/math-science/weather/climate-of-100-selected-us-cities` (visited on 12/18/2020).

[2] Hastie et al. *An Introduction to Statistical Learning with Applications in R (ISLR)*. 2013.