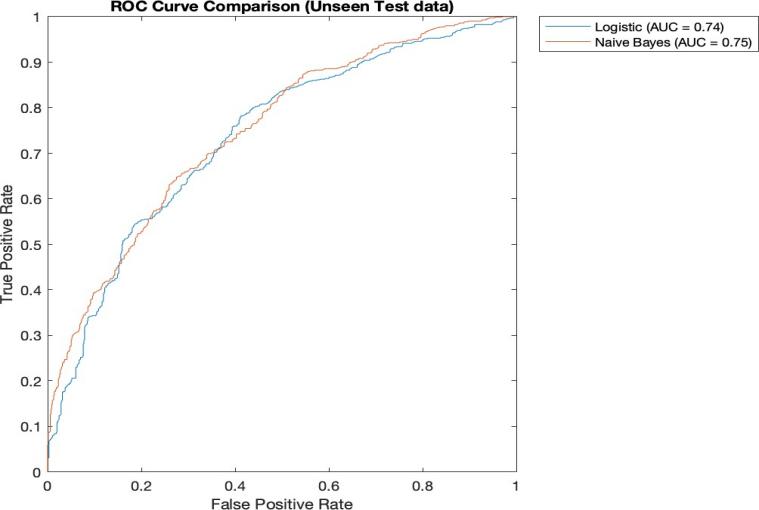
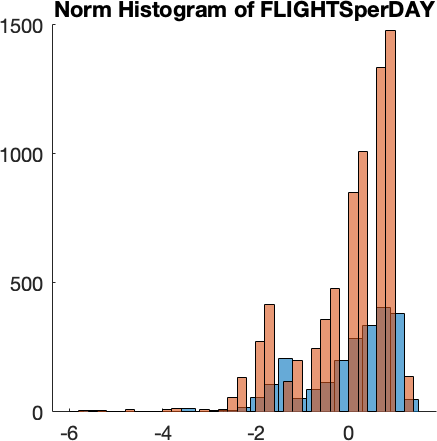
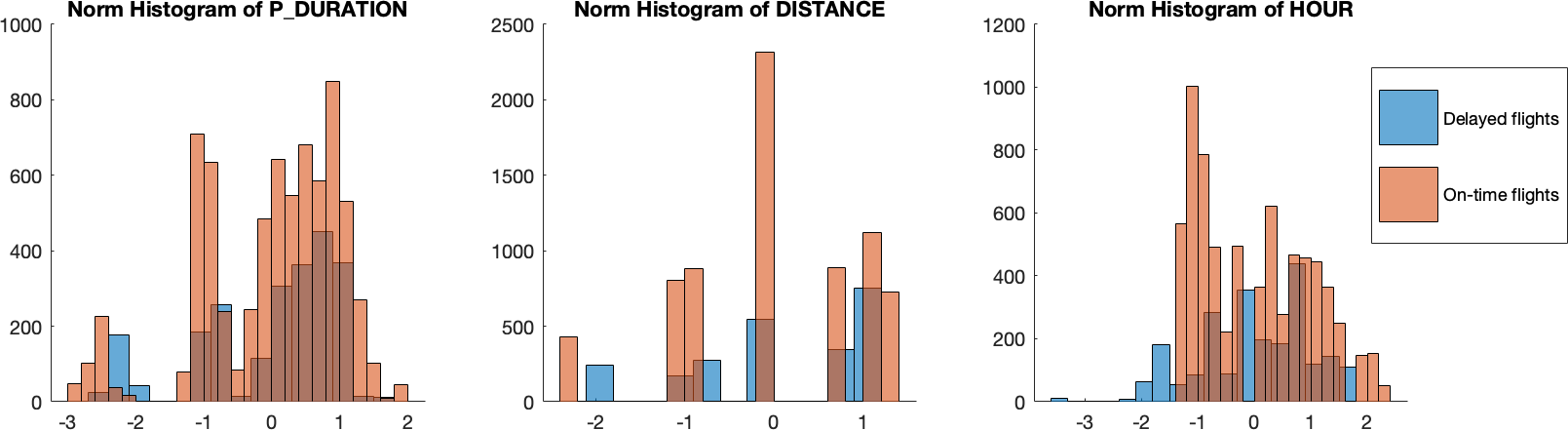
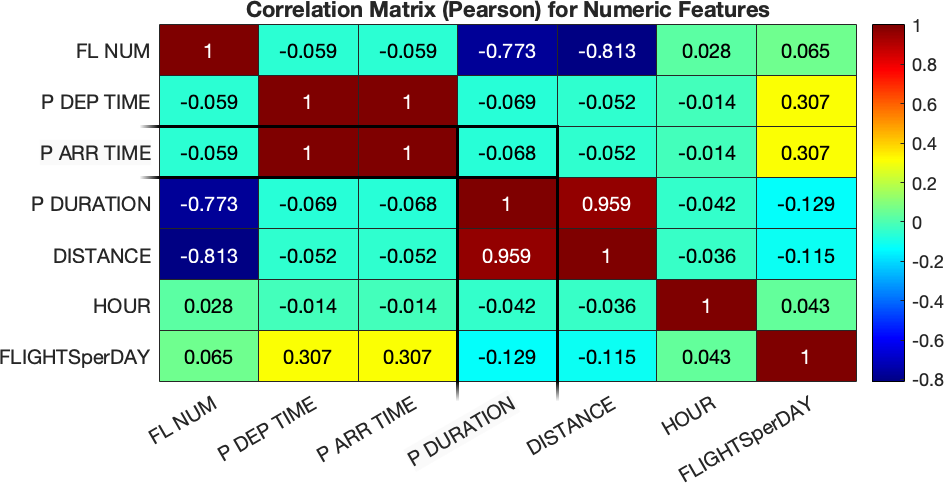
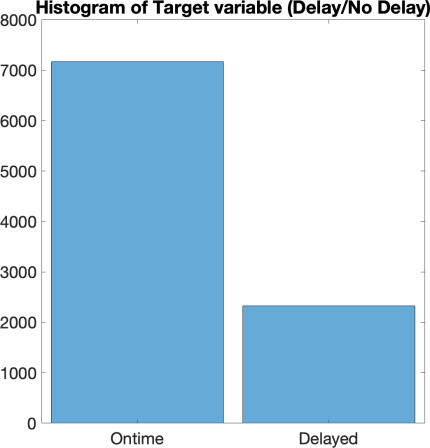
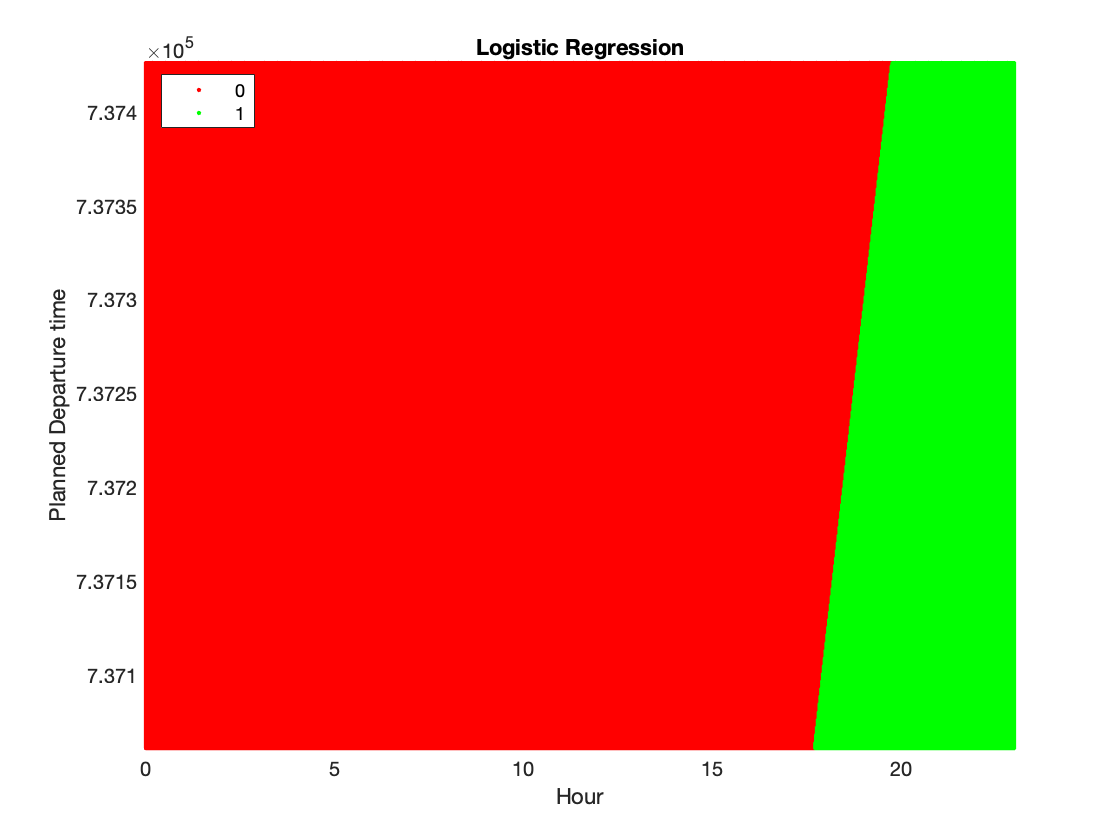
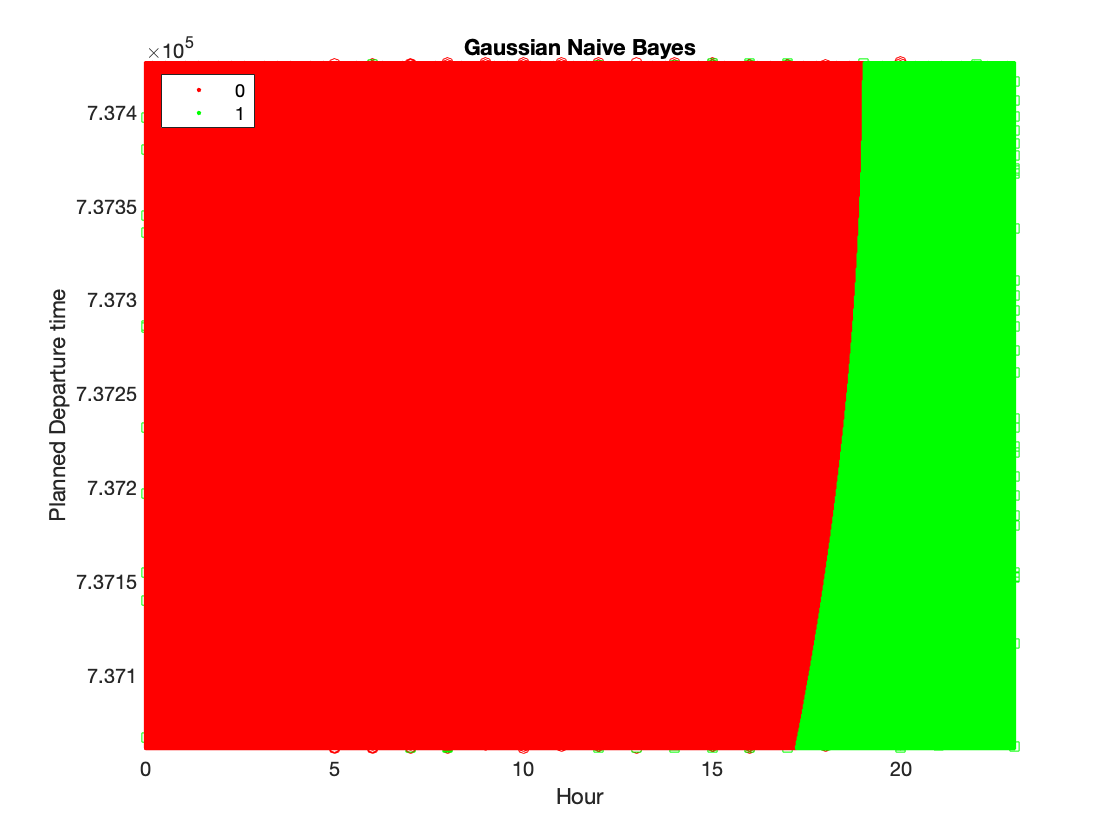
|  |  |
| --- | --- |
| **A comparison of Random Forest Regression and Logistic Regression (LR) on predicting flight delays** | |
| **Description and motivation**  We will solve the binary classification problem of predicting flight departure delays based on information only available before departure time at Westchester County airport (HPN) in New York using Logistic Regression and Naïve Bayes. We will aim to compare our results to those obtained by Scott Cole and Thomas Donoghue (2017)1 for Logistic Regression, then apply a Naïve Bayes algorithm to the same dataset and use performance metrics to assess model performance. | |
| **Exploratory Analysis**   * Dataset: FinTech App Usage Data from Kaggle * The dataset consists of US domestic flight details from the Bureau of Transportation Statistics. We focused on the 2018 data and HPN airport to aid in faster analysis and lower computation time. * The data was cleaned to only include flight details known at the time of departure. * Feature engineering and feature selection were carried out to include categorical predictors like month, hour and day to our model. Some of these steps are outlined in Scott Cole and Thomas Donoghue (2017). * A flight delay is defined as a flight departing 15 minutes or later than the planned departure time. * The cleaned dataset consists of 9495 rows, 12 features and 1 target column. There are 5 numeric, 5 categorical and 2 date-time features. * The target column contains a 1 if a flight was delayed and 0 otherwise. * A class imbalance problem exists as shown by the graphs below with 75% of flights on time and only 25% delayed. * The numeric features in our data do not follow normal distributions even after normalization. Furthermore, most of the numeric features of delayed flights are more negatively skewed as shown in the table. * The Pearson correlation coefficient shows a strong linear correlation between some features. We address this in later steps. * Analyzing the histograms of the categorical predictors shows JetBlue Airways had the greatest proportion of all flights and all delays. Most flights travelled from HPN airport to ATL airport, however the greatest proportion of delays were of flights going to PBI airport. Of all evening flights, more were delayed than on time. Flights on Monday or Friday had the most delays when compared to other days of the week.   ***Mean Standard Deviation Skew***  ***Delayed Ontime Delayed Ontime Delayed Ontime FL\_NUM*** *2938 3120 1816 1759 0.20 0.02*  ***P\_DURATION*** *148 147 33 29 -1 -0.88*  ***DISTANCE*** *800 785 277 242 -0.68 -0.49*  ***FLIGHTSperDAY*** *26 27 3.63 3.53 -1.33 -1.46*  ***Hour*** *15 11 5 4 -0.49 0.34* | |
| **Random Forest Regression** | **Naïve Bayes** |
| * Random Forest Regression is a supervised learning algorithm designed for regression tasks. * It operates by constructing multiple decision trees during training and averaging their outputs to improve predictions and reduce overfitting. * Unlike single decision trees, Random Forest uses an ensemble approach to achieve better accuracy and robustness. * The algorithm randomly selects subsets of data (with replacement) and features for tree building, which ensures diversity and reduces correlation between trees. * Random Forest can handle both numeric and categorical predictors effectively. * The final prediction for regression tasks is the average output of all decision trees in the forest. andIt includes built-in feature importance evaluation, enabling identification of the most impactful predictors..   **Advantages Disadvantages**   * + Reduces overfitting with averaging. ❗ May overfit small datasets   + Models non-linear relationships effectively.   + Identifies feature importance.   + Handles missing data and high-dimensional inputs. engineering and feature selection are important. problems.   + Requires minimal preparation; resistant to outliers ❗.     ❗ Computationally slow; less interpretable. | * Naïve bayes (NB) algorithm is called ‘naïve’ because it assumes that given any target class, the features are independent. In our example we have 2 classes (delayed or not delayed). * Generative classifiers like Naïve Bayes, learn a model of the joint probability *p( x , y )* of the inputs x and the labels y, and make their predictions by using Bayes rules to calculate *p( y l x )* and then pick the most likely label y.6 * A decision rule is used to pick the label, e.g., maximum a posteriori which selects the most probable hypothesis. * Usually, Naïve Bayes models perform surprisingly well despite the assumption of independence. * Assumes prior probabilities from class distribution in the training set, multinomial distribution for discrete features and gaussian distribution for continuous features but these can be adjusted to find the optimal model.   **Advantages Disadvantages**   * + Easy to implement. ❗ Very rare that all your predictors will be independent in   + Deals well with high dimensional data. real life scenarios. This can also lead to inaccurate   + Simple to run and efficient. posterior probability calculations as shown by Harry   + When the independence assumption is true it Zhang.8   + works better than most algorithms. a and irrelevant ❗ If an attribute in the test set was not present with a Robust to the presence of noise dat class in the training set, Naïve Bayes assigns zero   + attributes.4 probabilty7.. This means it will not be able to make a All probabilities needed to build the model are prediction for that instance.   + found in one scan, thus training is linear. ❗ Naïve Bayes works best with discrete features rather   7  Few hyperparameters. than continuous. |
| **Hypothesis Statement:**   * Both MLR and Random Forest models demonstrate limited predictive performance with similar RMSE values (MLR: 0.3717, RF: 0.3725) and low R2R^2R2 (~0.0015). * MLR indicates minimal significance for most predictors, suggesting weak relationships with the target variable. * These findings indicate weak or irrelevant signals in the dataset, limiting the models' effectiveness. * Naive Bayes will have a quadratic decision boundary whereas Logistic Regression will have a linear decision boundary. | **Methodology:**   1. Split data into a 70: 30 split for train and test data. The test data remains unseen to models until end. 2. 10 - fold cross validation on the ‘train’ data outlined above. From this we will obtain averages of AUC, accuracy, error, recall, and other performance metrics. This will provide information on the goodness of fit of models. The baseline model will use the features Scott Cole & Thomas Donoghue1 used. 3. Optimise models by feature selection, correcting target class imbalance and hyperparameter optimisation. 4. Evaluate which are optimal models based on performance metrics. 5. Predict the unseen test set outlined in step 1. Measure and compare predictive performance of the optimized models. Record train and test times. 6. We will try to minimize the training error as a proxy for minimizing the generalization error which in general is impossible to calculate exactly. |
| **Experimental results, parameter choices and feature selection**:  ***Multiple Linear Regression:***   * The initial model with 51 predictors showed weak performance due to irrelevant features. * Feature Selection Removing non-significant predictors didn't improve performance significantly.. * The best-performing model did not show substantial improvement despite feature selection, indicating that further data refinement and feature engineering are needed. The decrease in accuracy and performance metrics suggests that more meaningful predictors should be incorporated.   ***Random Forest Regression***  • The Random Forest model with 100 trees showed weak performance and no meaningful feature importance.   * Random Forest yielded zero importance for all predictors, showing weak signal in the data.   The best model was effectively similar to the baseline model, showing no significant improvement despite using the Random Forest algorithm. Given the absence of important predictors, further exploration of data and alternative models is required.l.  *The table above shows the Average AUC and classification error calculated after 10-fold cross validation on the training set.*  ***Final Model Comparison:***   * The best two models outlined above were used to predict the previously unseen test set. Results outlined below.   *The table above shows the Average AUC and Error calculated after 10- fold cross validation on the training set.*  ***Decision Boundary Comparison***   * NB produced a quadratic boundary in contrast to the linear boundary produced by LR. |
| **Analysis and Evaluation of results**:   * The average training AUC of the baseline model was higher than that achieved by Scott Cole & Thomas Donoghue (2017) 1. This could because we focused on one airport, or due to the random test-train split. It gave a better average training accuracy score than our final model, due to class imbalance . It was effectively trained to always return ‘not delayed’ as an output, so this needed improvement. By balancing the classes, misclassification errors increased by 10%, but the model was correctly identifying more delayed flights as precision and recall increased. * Feature selection further improve the model. We used Lasso regularisation rather than ridge regression because of the high dimensionality of the problem (it was important to remove redundant features). Lasso regularization eliminates coefficients if they do not contribute to the target variable, whereas Ridge regression shrinks the coefficients so the total number of features in model remains the same. 9 * After feature selection we obtained the final trained logistic model, results can be seen on the left. * We explored using univariate feature ranking through chi-squared tests for feature selection for NB. Here continuous variables are grouped into a specified number of bins. Higher Chi-squared values mean features are more dependent on the response variable. This yielded similar feature importance to those obtained from Lasso regularization. Thus, for comparability of the two final models only the features obtained from Lasso regularization were used. Feature selection reduced the average training AUC and average training accuracy post cross validation of the NB model. * For NB, changing the distributions of the numeric features from normal distributions to kernel density estimation improved the average training accuracy of the model. This is because numeric features like hour and planned departure time did not follow normal distributions. Through trial and error, I arrived at a +1-kernel width to maximise training accuracy. We now arrived at our final trained NB model. * We then compared how both our final models performed when predicting the unseen test set. NB training time was lower than LR as expected, however NB predict time was 4x higher than LR. This could be due to the way the MATLAB function works or it could be generally indicative of higher test times for NB, although I have not been able to find any literature to support the latter. * In our final model comparison, the test AUC of NB is marginally higher than that of LR (could be due to random test and train split). Average train error across the 10 folds was lower in NB and the test error was the same for both models. O. Kupervasser proved that Naive Bayes gives minimal mean error over uniform distribution of all possible correlation between characteristic variables.11 Thus proving why NB performs surprisingly well. * Variance refers to the amount by which the prediction of a model changes on a different train set. 10 Our baseline logistic model had too many features, and high complexity thus it was prone to overfitting and high variance. In contrast, our final models have very little difference in test and train errors individually, suggesting they have low variance. * Bias refers to the error introduced into the model by using simpler models. 10 Our final models are simple due to NB assumption of independence and the reduction in number of features. Thus, our final models will suffer from high bias. |
| **Lessons Learned**   * Optimizing for both models involves feature engineering and feature selection. However, in NB, tuning parameters like prior probabilities and distributions of numeric features can help reach optimal model performance. * Target class imbalance leads to inefficient model performance in terms of precision and recall.   **Future Work:**   * Other methods to correct target imbalance e.g., SMOTE (Synthetic Minority Oversampling Technique) * Using MRMR (Max- redundancy Minimum relevance) for feature selection in Naïve Bayes. * Explore effects of using variants of NB classifier.7 * For Logistic regression model, work on feature engineering perhaps by adding weather information which other papers have done. Could try to use Ridge regression for feature selection instead of Lasso. * Adding weather information to improve both models as this has been done in other papers. * Using data from more years, as we only used 2018 data would improve our model. |
|  |

References:

1.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **LR Baseline** | **NB baseline** | **LR Baseline (class balanced)** | **NB baseline (class balanced)** |
| **Avg Train AUC** | **0.76** | **0.74** | **0.74** | **0.73** |
| **Avg Train Error** | **0.221** | **0.232** | **0.322** | **0.323** |

|  |  |  |
| --- | --- | --- |
|  | **Naïve Bayes** | **Logistic Regression** |
| **Avg Train AUC** | **0.76** | **0.75** |
| **Test AUC** | **0.75** | **0.74** |
| **Avg Train Error** | **0.316** | **0.323** |
| **Test Error** | **0.325** | **0.325** |
| **Train Time** | **0.015** | **0.032** |
| **Predict time** | **0.018** | **0.003** |

1 Cole. S. and Donoghuge. T, “*Predicting departure delays of US domestic flight*s “, 2017.

2 Hastie. T, Tibshirani. R. and Friedman. J, *“ The Elements of Statistical Learning, Data mining, Inference, Prediction “,* 2008, Chapter 4, *page 110*.

3Murphy. K, “ *Machine Learning A Probabilistic Perspective “*, 2012.

4 Moereira.J, Carvalho.A and Horvath.T, “*A general introduction to Data Analytics "*, 2019, Chapter 9, page 207.

5 Tuffery.S, “*Data Mining and Statistics for Decision Making “,* First Edition, 2011, Chapter 11, page 478.

6 Andrew Y. Ng , Michael I. Jordan*, “On Discriminative vs. Generative classifiers: A comparison of logistic regression and Naive Bayes "*, 2001, <https://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf>(accessed 6th December 2020).

7 K. M. Al-Aidaroos, A. A. Bakar and Z. Othman, "Naïve bayes variants in classification learning”, 2010, pp. 276-281, <https://ieeexplore.ieee.org/document/5466902> (accessed 6th December 2020).

8 Zhang. H, “ *EXPLORING CONDITIONS FOR THE OPTIMALITY OF NAÏVE BAYES ”* , 2005, <http://www.yaroslavvb.com/papers/zhang-exploring.pdf> (accessed 6th December 2020).

9 Statistics How To, *Ridge Regression, <https://www.statisticshowto.com/ridge-regression/> (*accessed 6th December 2020).

10 [Ani Karenova K](https://medium.com/%40karenovna.ak/part-ii-evaluating-a-predictive-model-cross-validation-and-bias-and-variance-tradeoff-9874b836cd2e)*[, ”Part II. Model Evaluation: Cross Validation, Bias and Variance Tradeoff and How to Diagnose Overfitting“ ,14/04/2019, https://medium.com/@karenovna.ak/part-ii-evaluating-a-predictive-model-cross-validation-and-bias- and-variance-tradeoff-9874b836cd2e (accessed 6th December 2020).](https://medium.com/%40karenovna.ak/part-ii-evaluating-a-predictive-model-cross-validation-and-bias-and-variance-tradeoff-9874b836cd2e)*

11 O.Kupervasser, ”The mysterious optimality of Naïve Bayes : Estimation of the Probability in the system of ‘classifiers’“, *Pattern Recognit. Image Anal.* **24,** 1–10 (2014), [https://link.springer.com/article/10.1134/S1054661814010088#citeas](https://link.springer.com/article/10.1134/S1054661814010088) (accessed 6th December 2020).