PROJECT REPORT*

SENTIMENT ANALYSIS ON CUSTOMER REVIEWS OF A PHOTO COMPANY

GOKAY BULUT

*Note: Since the company owns all intellectual property rights, this is not the original presentation provided to the company. Rather, it is a report explaining my work.

AGENDA

- Introduction
- Data Wrangling
- Exploratory Data Analysis (EDA)
- Modeling
- Conclusion

INTRODUCTION

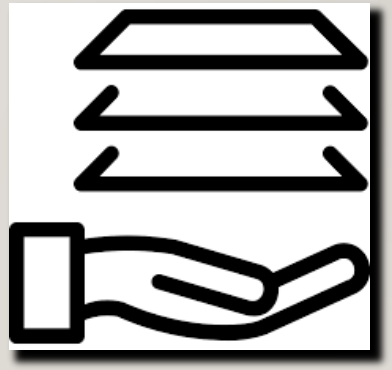
Problem:

- capture the most used words by customers,
- build a sentiment analysis model that predicts whether a customer liked a product or not, based on the reviews.

INTRODUCTION

Data set:

- 5157 rows of customers' reviews and ratings,
- provided by the Company.



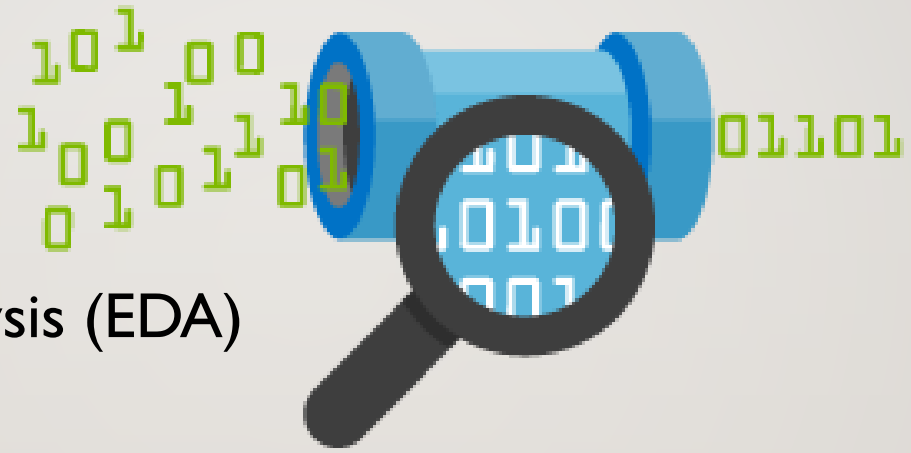
INTRODUCTION

Features:

- name - name of the reviewer (string)
- lastname - lastname of the reviewer (string)
- sex - sex of the reviewer (string)
- rating - rating given by the reviewer for the type of product / service (float)
- product - type of product / service (string)
- review - the text of the review (string)
- review_date - date the review was provided (string)
- reviewer_id - unique number of the reviewer (integer)

AGENDA

- Introduction
- Data Wrangling
- Exploratory Data Analysis (EDA)
- Modeling
- Conclusion

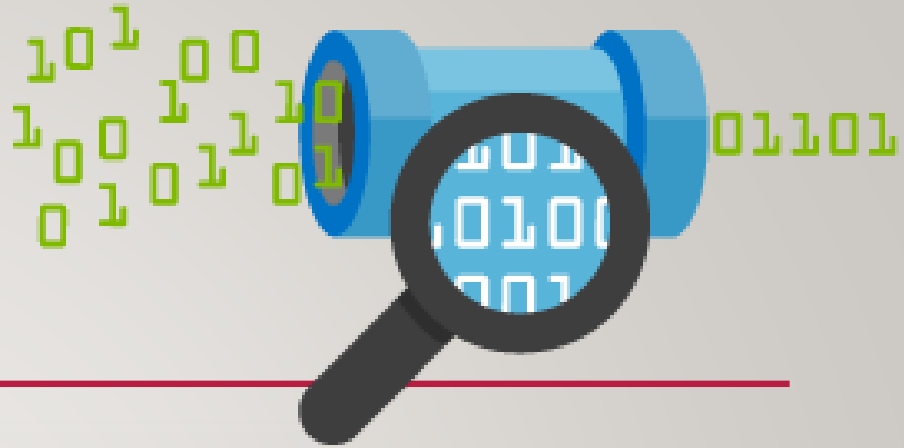


DATA WRANGLING

	sex	rating	product	review	review_date	reviewer_id		clean_text
0	f	5.0	service	İstanbul'da en sevdiğim mekan, analog dostu. M...	2017-03-08	44720	istanbulda	g dostu minicik...
1	f	5.0	studio	KESİNLİKLE ÇOK GÜZEL BİR STÜDYO	2017-10-17	46945	kesinl	ok guzel bir studyo
2	f	5.0	service	Burayı hep sevdim	2017-01-03	92805		buray sevdim
3	m	5.0	wedding	Düğün fotoğrafı için gitmiştik, çok güzel çeki...	2017-01-19	51670	dugun fotograf icin gitmistik cok guzel cekiml...	
4	f	2.0	service	O kadar iyi diil tabi. Isim var sadece	2017-11-05	79719		kadar iyi diil tabi isim var sadece

- column names → to snake_case naming convention,
- dropped redundant columns & rows, and rows with missing values,

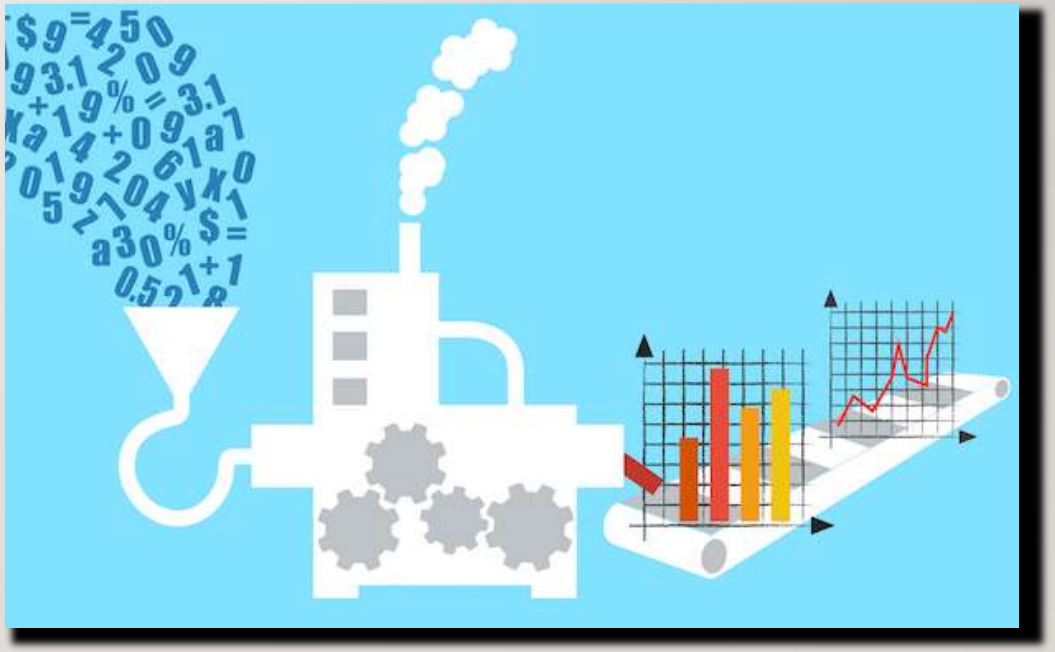
DATA WRANGLING



- 'sex' and 'product' columns → categorical type for efficient computing,
- 'review_date' column → datetime type for better analysis,
- preprocessed text for machine learning by
  - stripping the html tags,
  - converting the text to lower_case,
  - lemmatizing the text,
  - removing extra lines, accented or special characters, digits and stopwords.

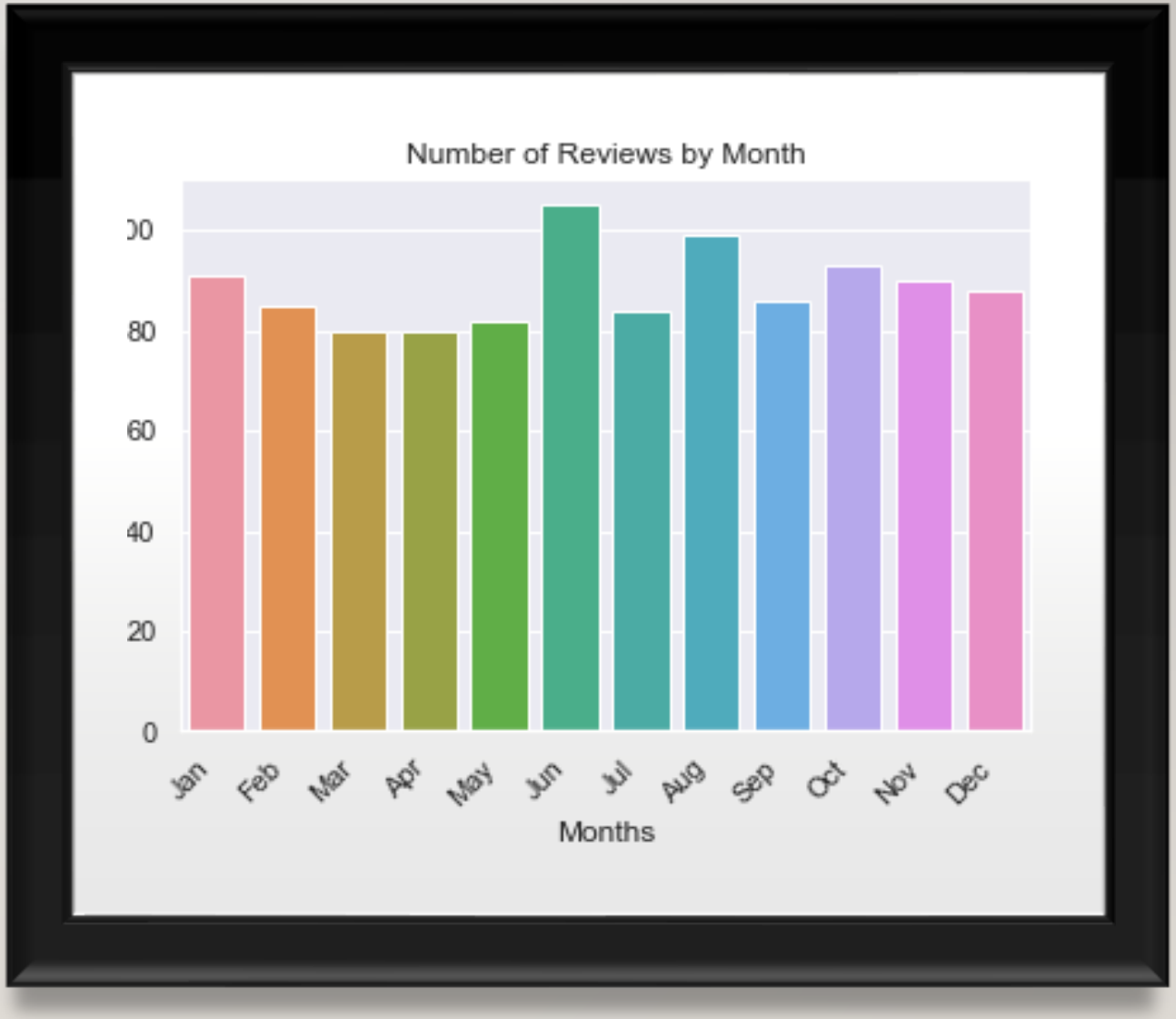
AGENDA

- Introduction
- Data Wrangling
- Exploratory Data Analysis (EDA)
- Modeling
- Conclusion



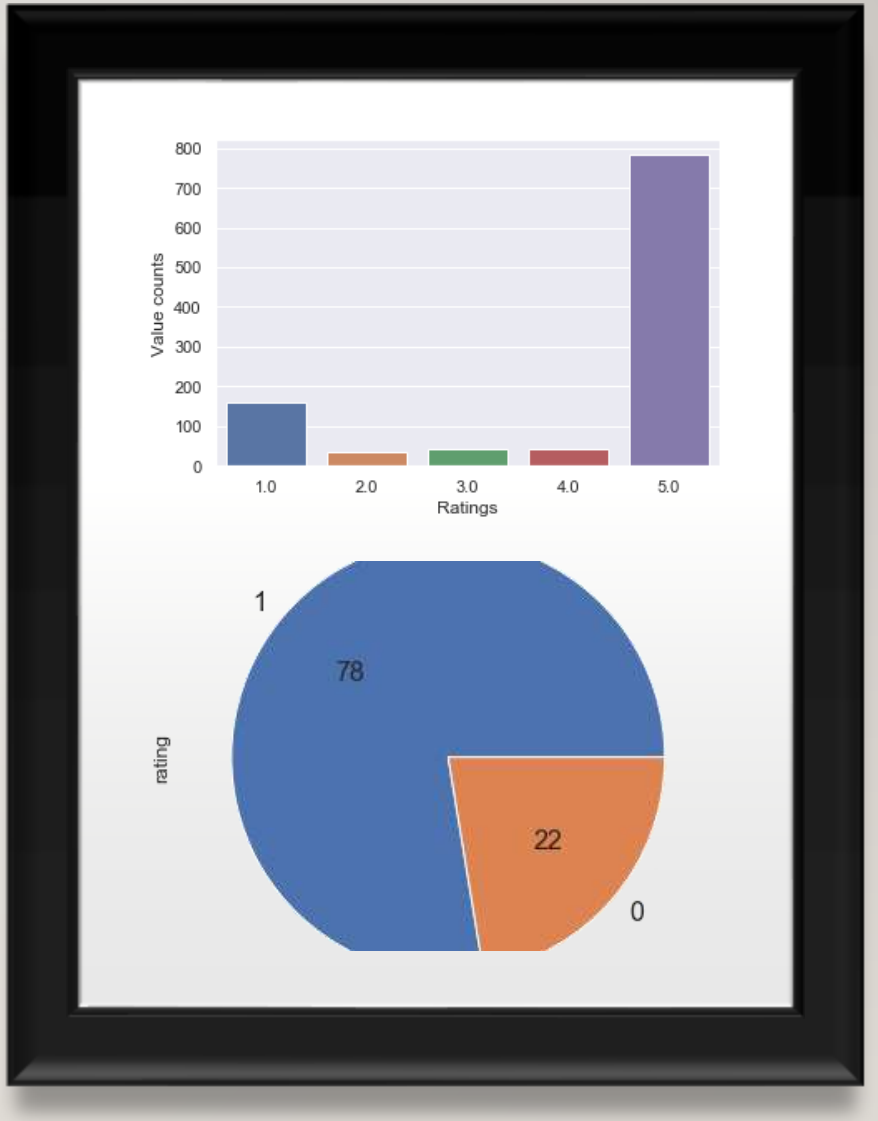
EDA
"REVIEW_DATE" FEATURE

- spans Jan 1st - Dec 31st, 2017,
- # reviews decrease towards spring and increase in summer,



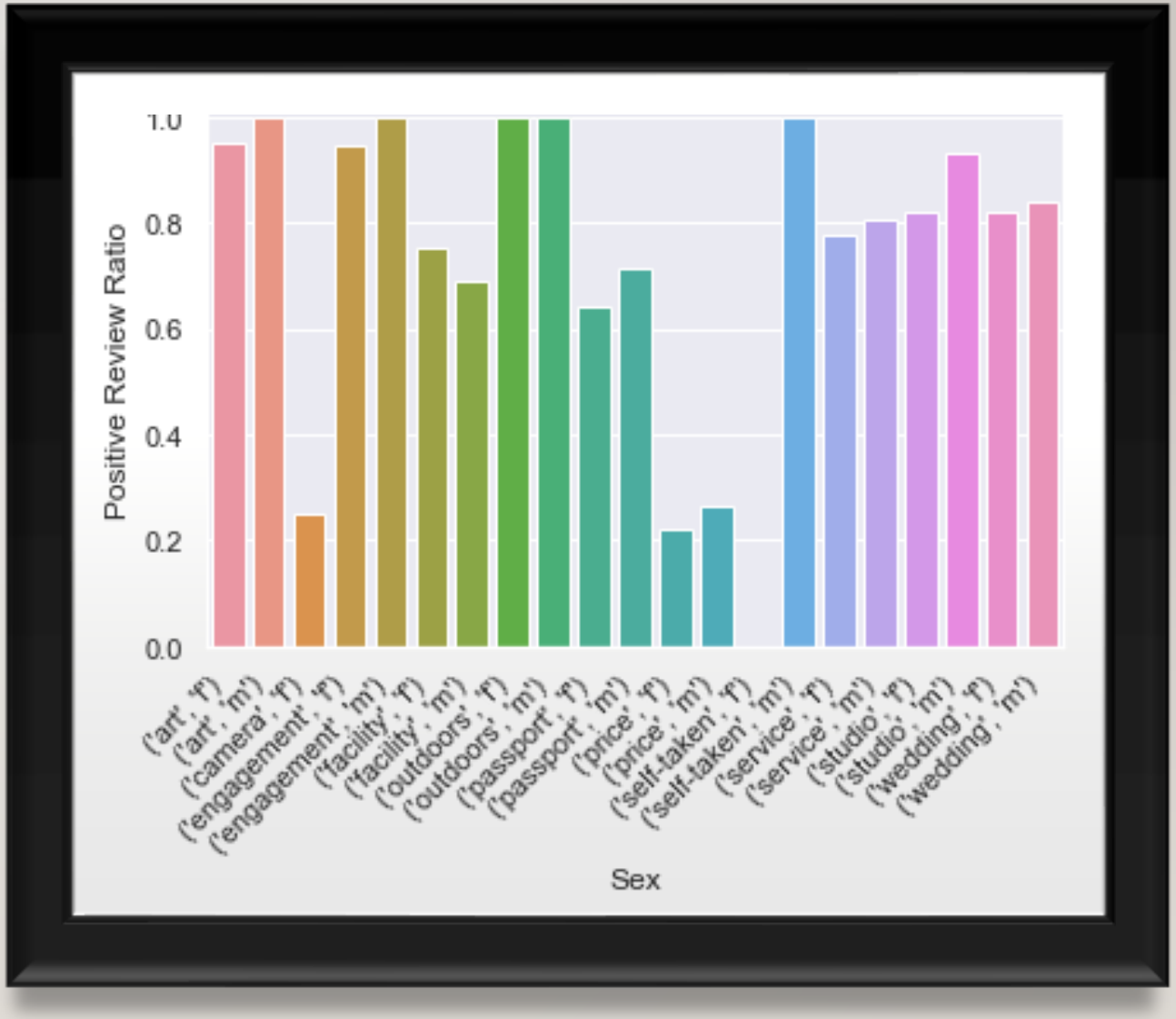
EDA
"RATING" FEATURE

- Classes highly imbalanced
- Categorized 'rating' as 0 and 1 to reduce imbalance,
- 78% of reviews positive,



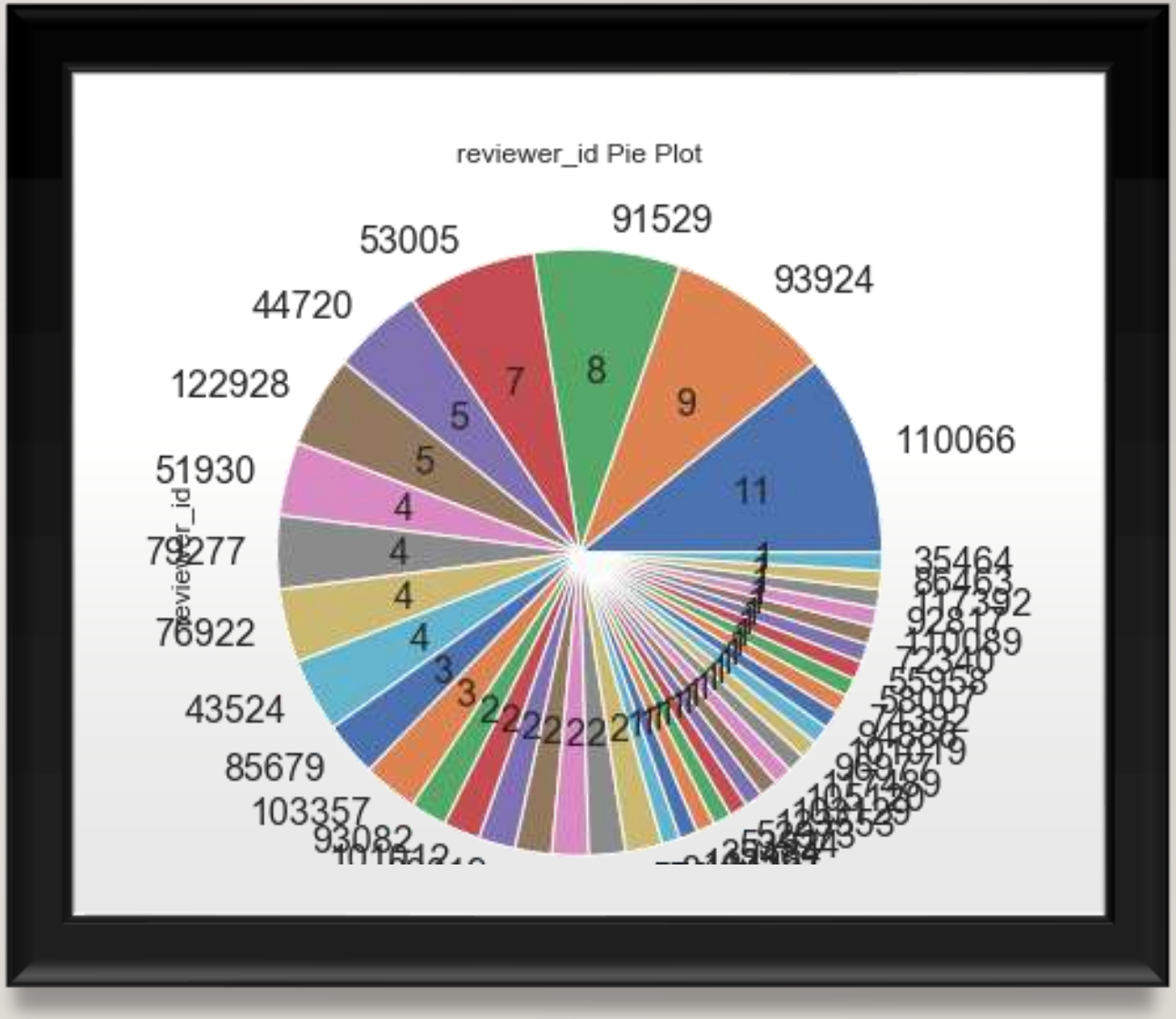
EDA
"SEX" FEATURE

- 51% of reviewers male,
- Ratio of positive reviews does not significantly differ by gender,



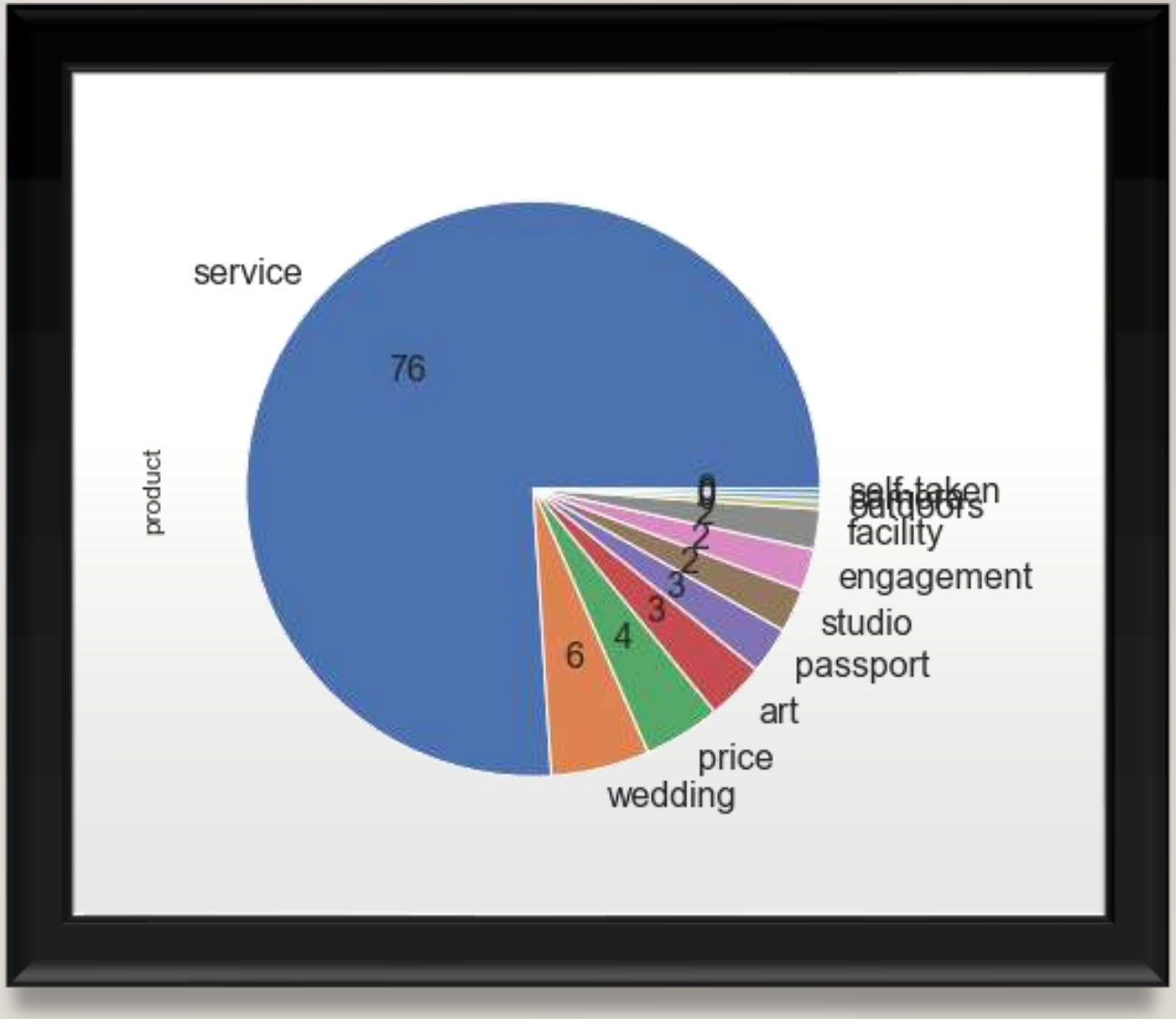
EDA
"REVIEWER_ID" FEATURE

- Some reviewers write more reviews.
- Beneficial to keep track of these customers.



EDA
"PRODUCT" FEATURE

- 76% of reviews are on service provided, i.e. customer satisfaction,
- wedding photos and prices also reviewed more,



EDA
"PRODUCT" FEATURE

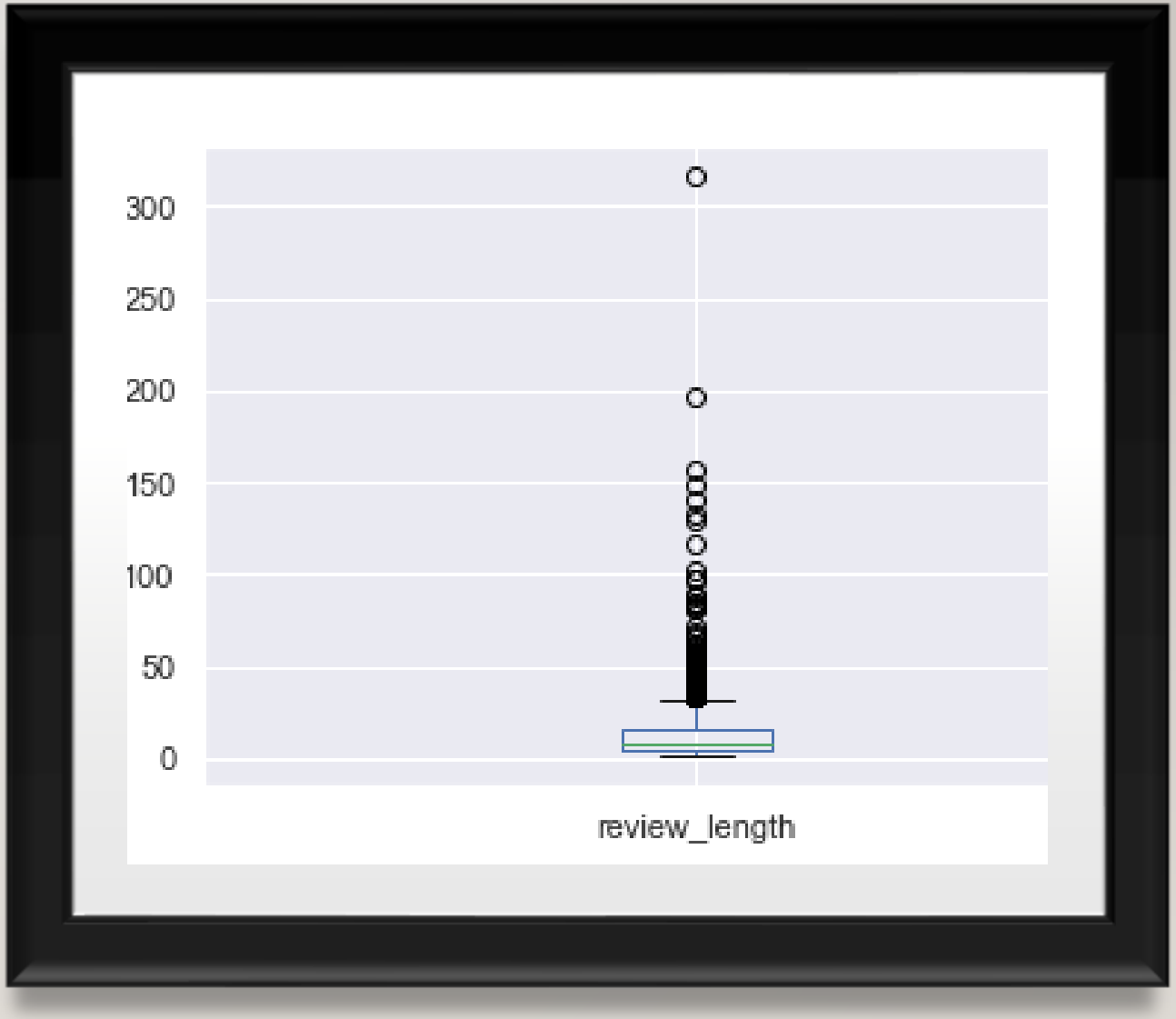
- 'price' = lowest rating (reviewed only 4%),
- 'outdoors' photo shoots = highest rating (reviewed only 0.5%),

EDA
"PRODUCT" FEATURE

- Price review highs might indicate price changes in July and December,

EDA
"REVIEW" FEATURE

- min 1,
- max 317,
- average 14 words long,
- Negative reviews longer,
- Big word share between positive and negative reviews,



EDA
"REVIEW" FEATURE

- beautiful,
- good,
- quality,
- recommend,
- thanks, professional, smiling, glad and fast = most used positive

EDA

"REVIEW" FEATURE

- bad,
- expensive,
- never,
- prices,
- place, absolutely, money, even and wedding = most used negative.

AGENDA

- Introduction
- Data Wrangling
- Exploratory Data Analysis (EDA)
- Modeling
- Conclusion

MODELING: MACHINE LEARNING

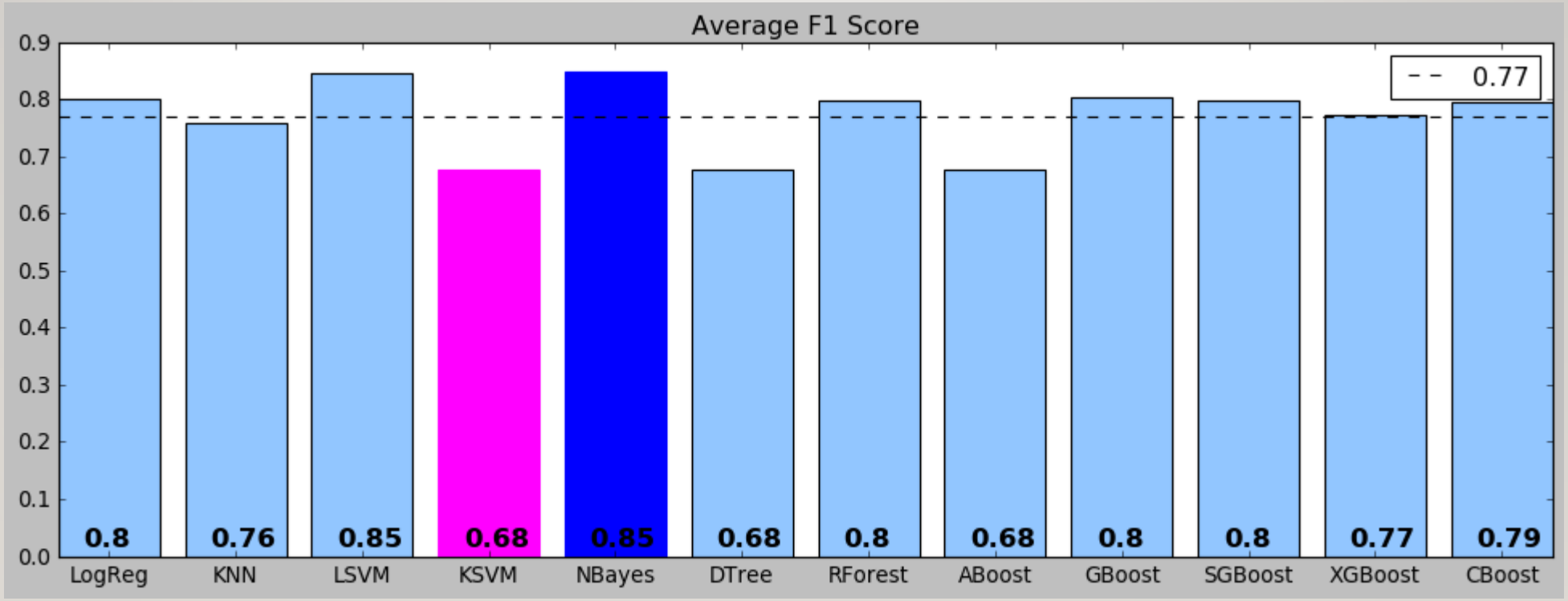
12 ML ALGORITHMS:

- Logistic Regression,
- K-Nearest Neighbors,
- Linear SVM,
- Kernel SVM,
- Naive Bayes,
- Decision Trees,
- Random Forest,
- AdaBoost,
- Gradient Boosting,
- Stochastic Gradient Boosting,
- XGBoost,
- CatBoost

7 BOW METHODS:

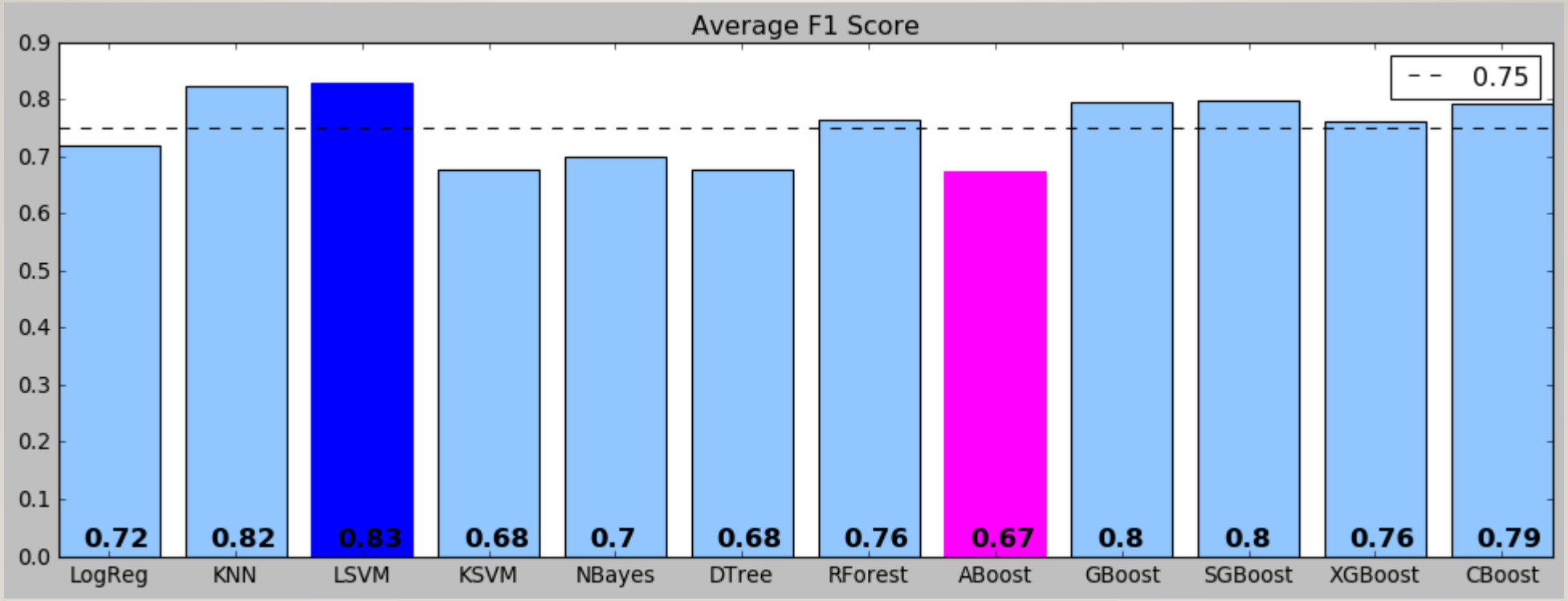
- Count Vectorizer
- Tfldf Vectorizer
- Hashing Vectorizer
- SMOTE
- PCA with SMOTE
- Truncated SVD with SMOTE
- Word2Vec

MODELING WITH MACHINE LEARNING
COUNT VECTORIZER



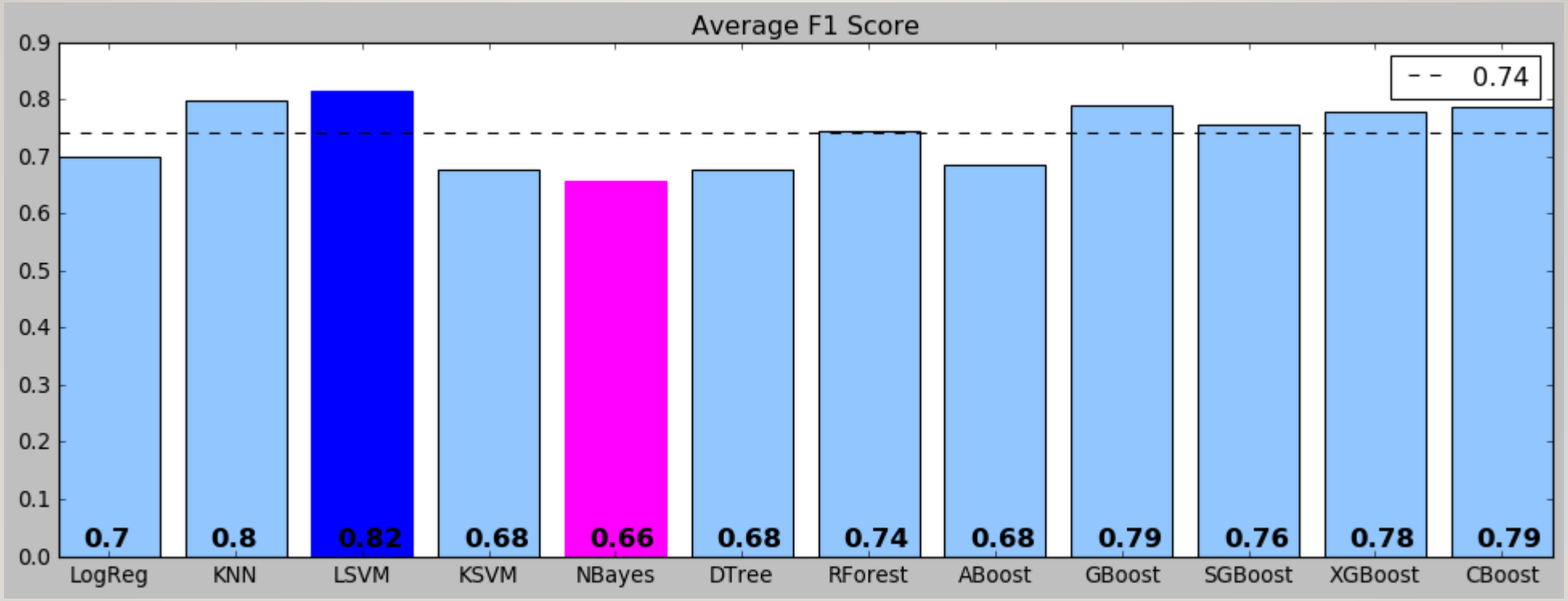
MODELING WITH MACHINE LEARNING

TF IDF VECTORIZER



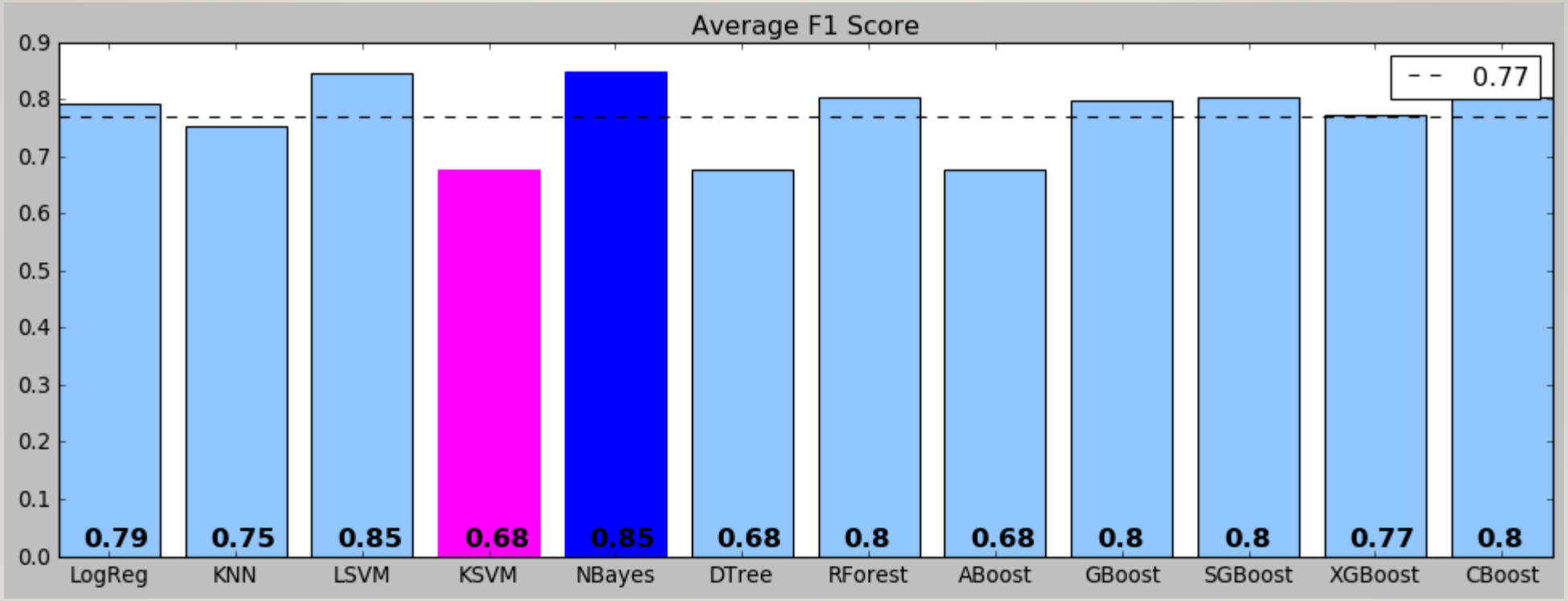
MODELING WITH MACHINE LEARNING

HASHING VECTORIZER



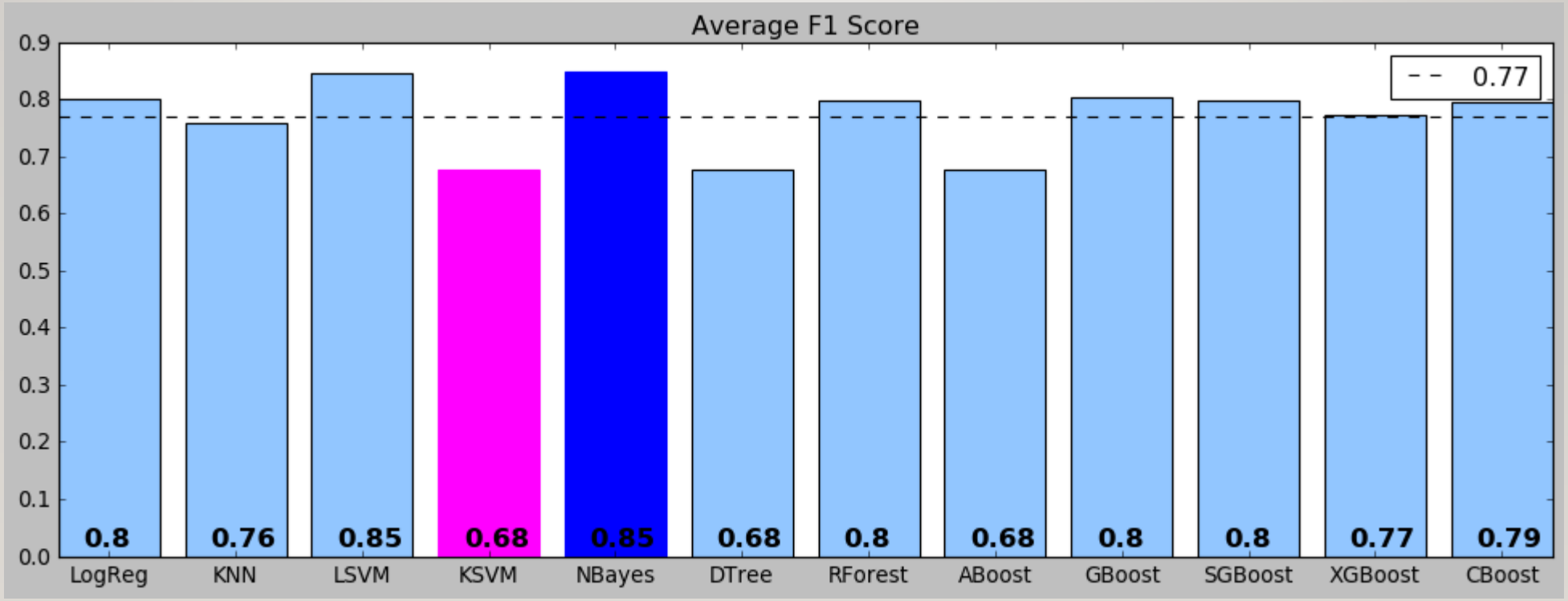
MODELING WITH MACHINE LEARNING

COUNT VECTORIZER AND SMOTE



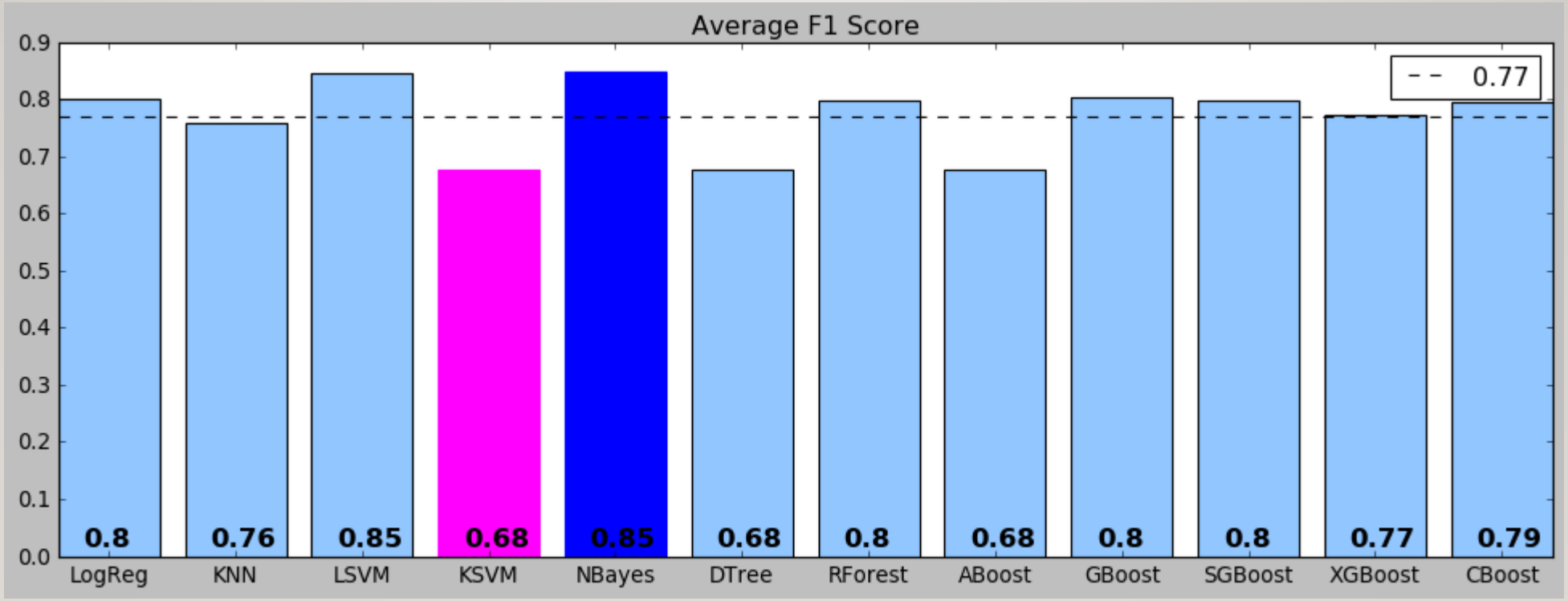
MODELING WITH MACHINE LEARNING

COUNT VECTORIZER AND PCA + SMOTE



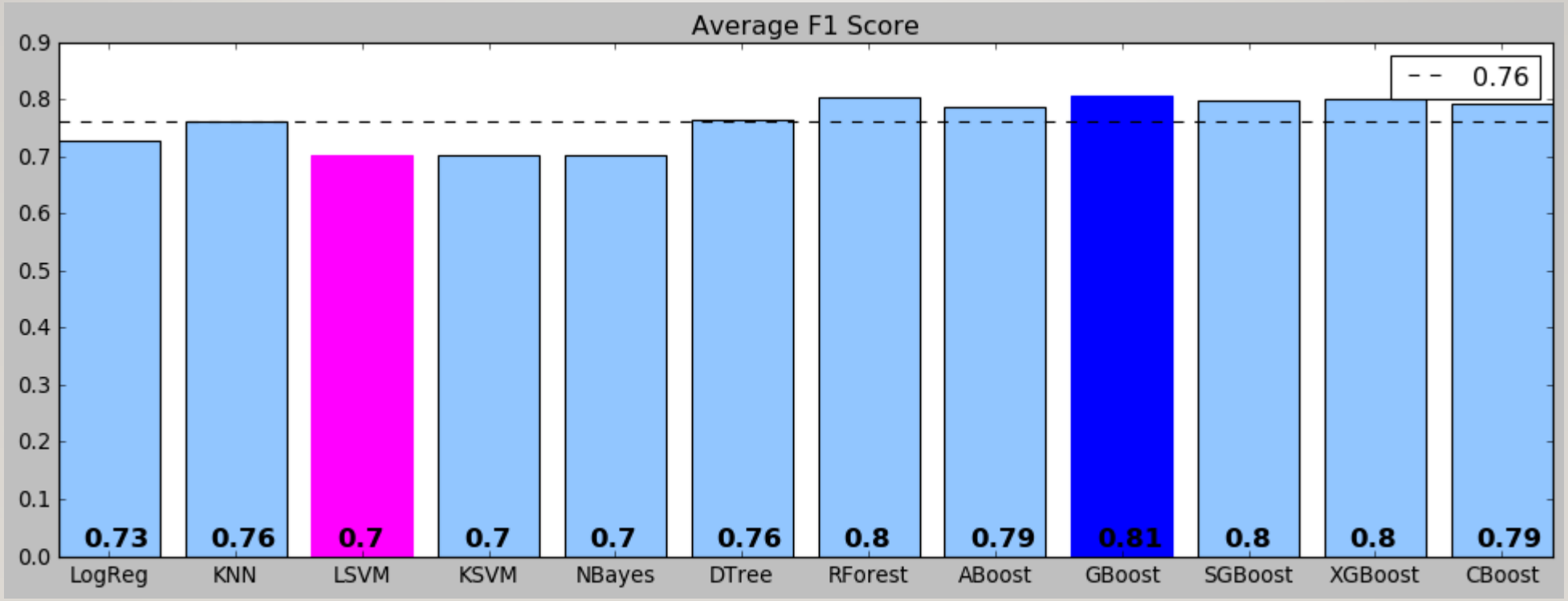
MODELING WITH MACHINE LEARNING

COUNT VECTORIZER AND TRUNCATED SVD + SMOTE



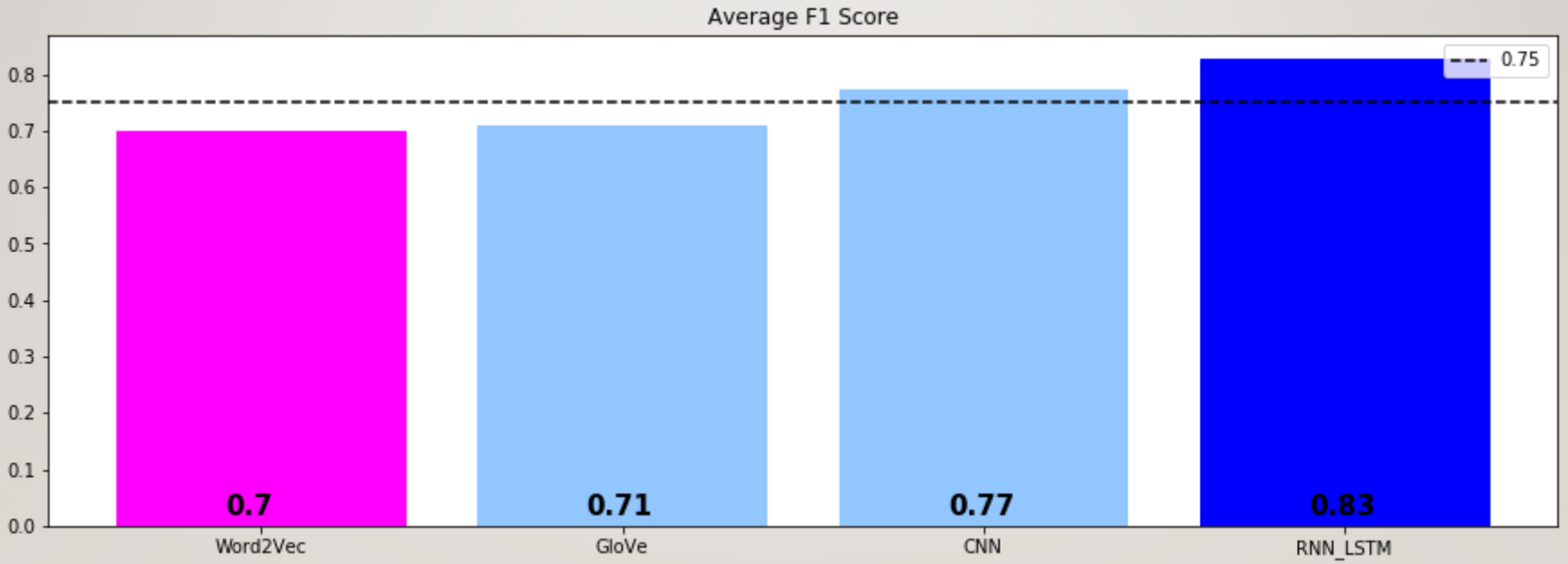
MODELING WITH MACHINE LEARNING

WORD2VEC



MODELING WITH DEEP LEARNING

WORD2VEC, GLOVE, CNN, RNN LSTM

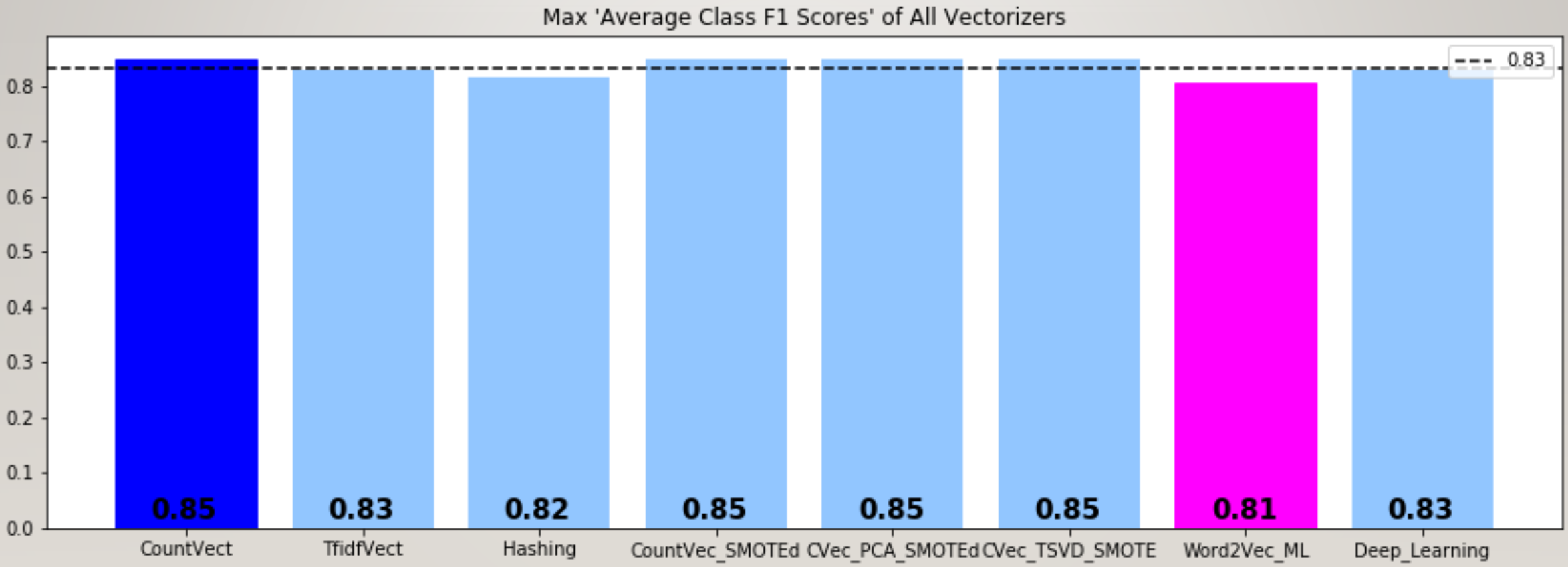


AGENDA

- Introduction
- Data Wrangling
- Exploratory Data Analysis (EDA)
- Modeling
- Conclusion

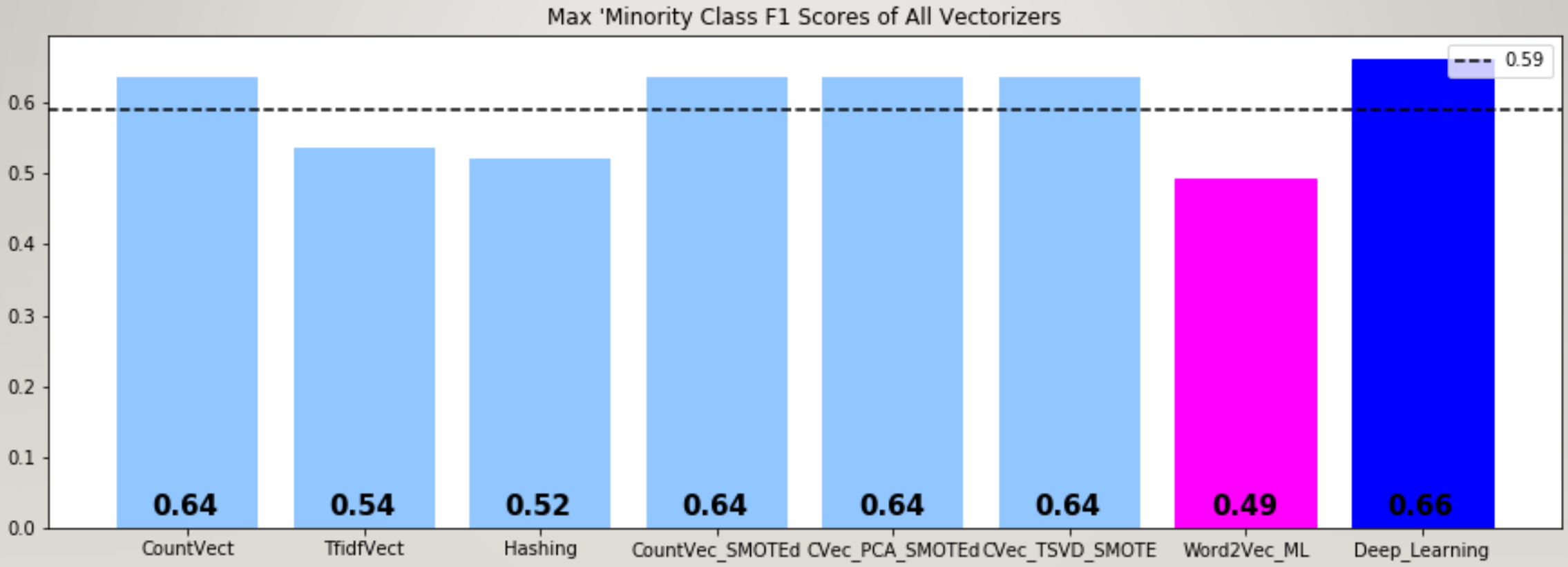
CONCLUSION

AVERAGE F-1 SCORES



CONCLUSION

MINOR F-1 SCORES



CONCLUSION
FINDINGS

- Overall, the best average F1 score of 0.85 was made by Count Vectorizer using Naive Bayes machine learning algorithm.
- Count Vectorizer + SMOTE, Count Vectorizer + PCA + SMOTE and Count Vectorizer + Truncated SVD + SMOTE using Naive Bayes also share the same best average F1 score of 0.85.

CONCLUSION
FINDINGS

- The best minority class F1 score of 0.66 was made by Long Short Term Memory (LSTM) Recurrent Neural Network (RNN).
- Word2Vec with machine learning has the lowest average score of 0.81 and minority score of 0.49.

CONCLUSION
FINDINGS

- The scores were negatively affected by 2 factors:
  - imbalance in data,
  - high rate of matching words among the classes,
- Possible areas for further improvement:
  - Need more data to train neural networks,
  - Implement Dask library for parallel processing to decrease run time.