EARNINGS PREDICTION FOR COLLEGE GRADUATES WITH MACHINE LEARNING

Nicholas Lemoff, Trigger Nandhra, Shuhan Luo, Kevin Ryu

INTRODUCTION

Research Question: How can we predict post-graduation income based on college majors?

Context and Motivation: Identifying key characteristics to financial success after college

Stakeholders: Students, parents and families, employers, counselors, educational institutions, policymakers

DATASET

174 row x 21 column dataset by the American Community Survey and FiveThirtyEight.

Each record represents a unique college major and its characteristics such as unemployment rate and major category.

```
1 df
```

	Rank int64 1 - 173	Major_code int64 1100 - 6403	Major object PETROLEUM... 0.6% MINING AND... 0.6% 171 others 98.8%	Total float64 124.0 - 393735.0	Men float64 119.0 - 173809.0	Women float64 0.0 - 307087.0
0	1	2419	PETROLEUM ENG...	2339	2057	282
1	2	2416	MINING AND MIN...	756	679	77
2	3	2415	METALLURGICAL...	856	725	131
3	4	2417	NAVAL ARCHITE...	1258	1123	135
4	5	2405	CHEMICAL ENGI...	32260	21239	11021
5	6	2418	NUCLEAR ENGIN...	2573	2200	373
6	7	6202	ACTUARIAL SCIE...	3777	2110	1667
7	8	5001	ASTRONOMY AN...	1792	832	960
8	9	2414	MECHANICAL EN...	91227	80320	10907
9	10	2408	ELECTRICAL ENG...	81527	65511	16016

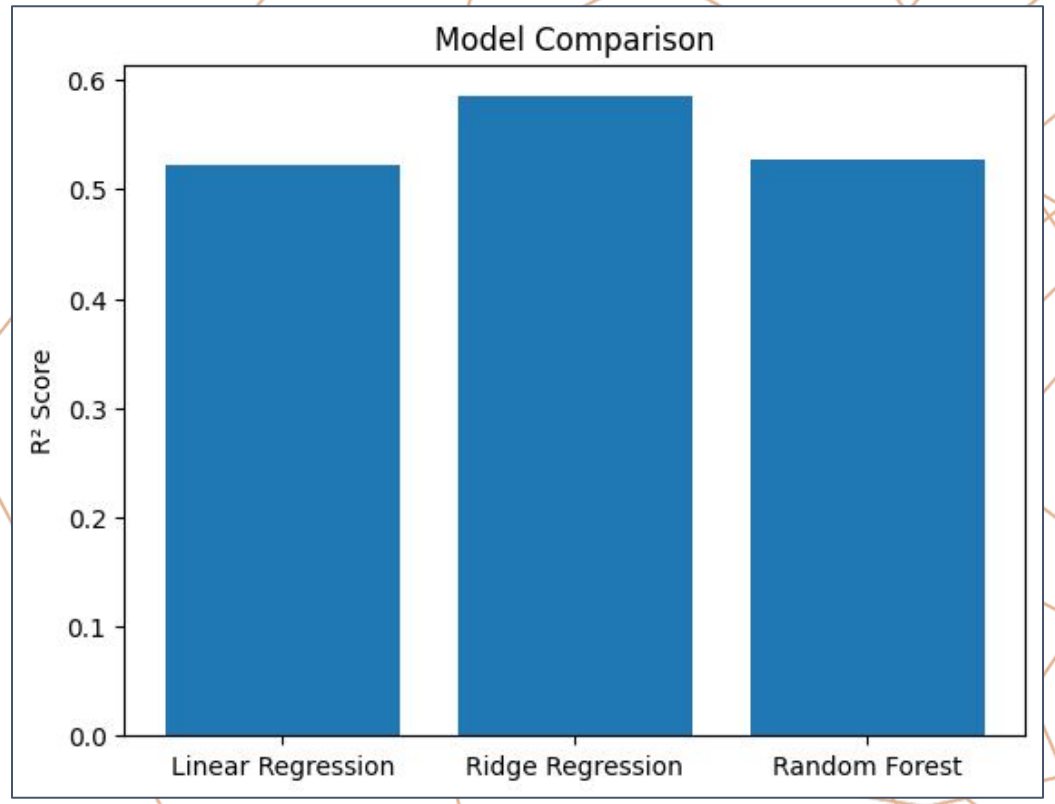
173 rows, showing 10 per page Page 1 of 18

```
1 df.shape
```

```
(173, 21)
```

METHODOLOGY

The project utilized random forest, ridge regression, and linear regression,



PREPROCESSING

- Read the dataset
- Dropped irrelevant columns
- Categorized features
- Transformed features

```
file_path = '/work/recent-grads.csv'
df = pd.read_csv(file_path)

X = df.drop(columns=['Median', 'P25th', 'P75th', 'Rank'])  # Excluding 'Median', 'P25th', 'P75th', 'Rank'
y = df['Median']

categorical_features = ['Major', 'Major_category']  # Update with actual categorical columns
numerical_features = X.select_dtypes(include=['int64', 'float64']).columns.tolist()

categorical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='most_frequent')),
    ('onehot', OneHotEncoder(handle_unknown='ignore'))
])

numerical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='mean')),
    ('scaler', StandardScaler())
])

preprocessor = ColumnTransformer(
    transformers=[
        ('num', numerical_transformer, numerical_features),
        ('cat', categorical_transformer, categorical_features)
    ]
)
```

MODELS

Linear Regression, Ridge Regression, and Random Forest for modeling.

```
linear_model = LinearRegression()
ridge_model = Ridge(alpha=10, solver='auto', random_state=42)
random_forest_model = RandomForestRegressor(
    n_estimators=100,
    max_depth=20,
    min_samples_split=2,
    min_samples_leaf=1,
    max_features='log2',
    random_state=42
)
```

ENSEMBLE

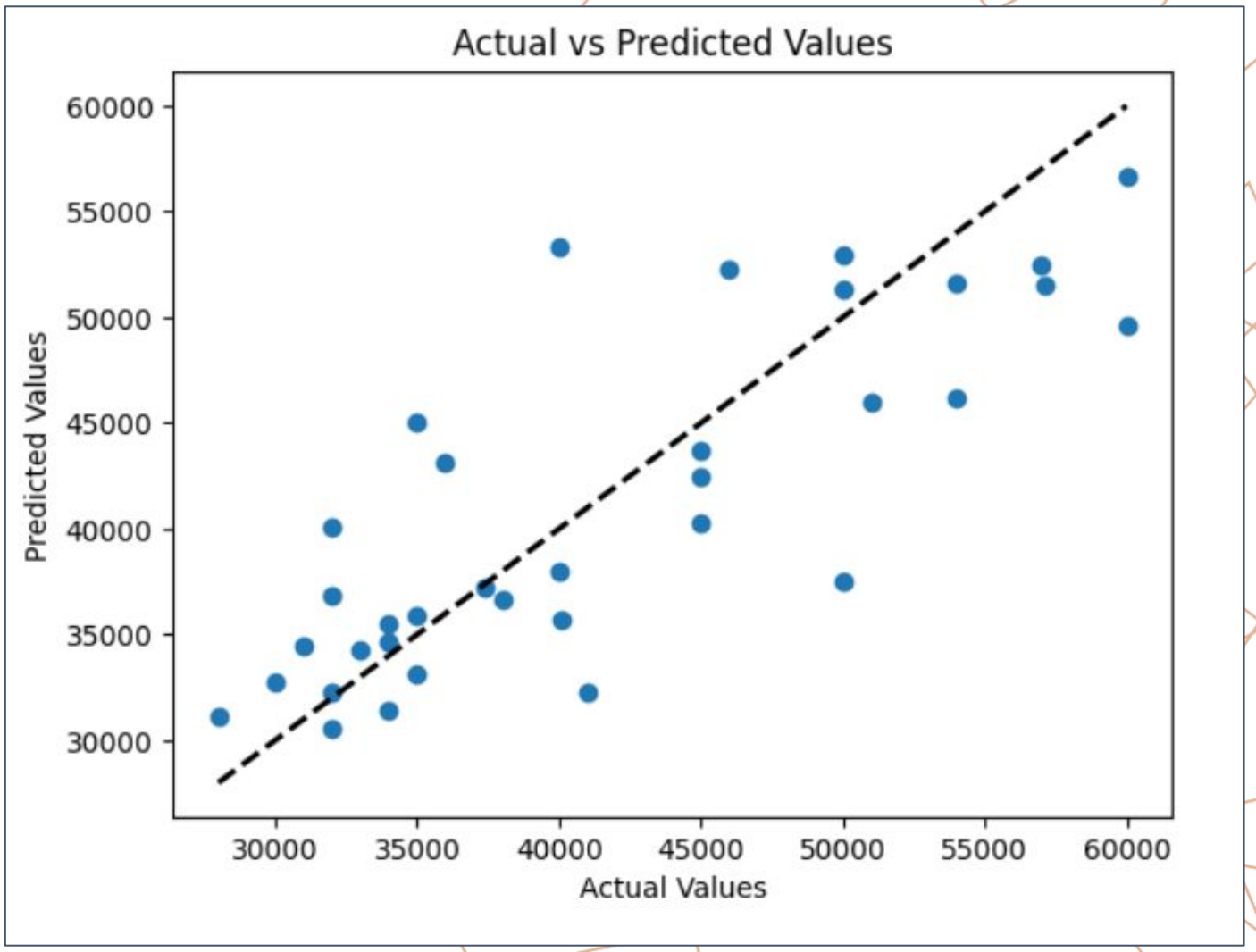
Establishing the ensemble and pipeline

```
ensemble = VotingRegressor(
    estimators=[
        ('lr', linear_model),
        ('ridge', ridge_model),
        ('rf', random_forest_model)
    ]
)

pipeline = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('regressor', ensemble)
])
```

ACCURACY

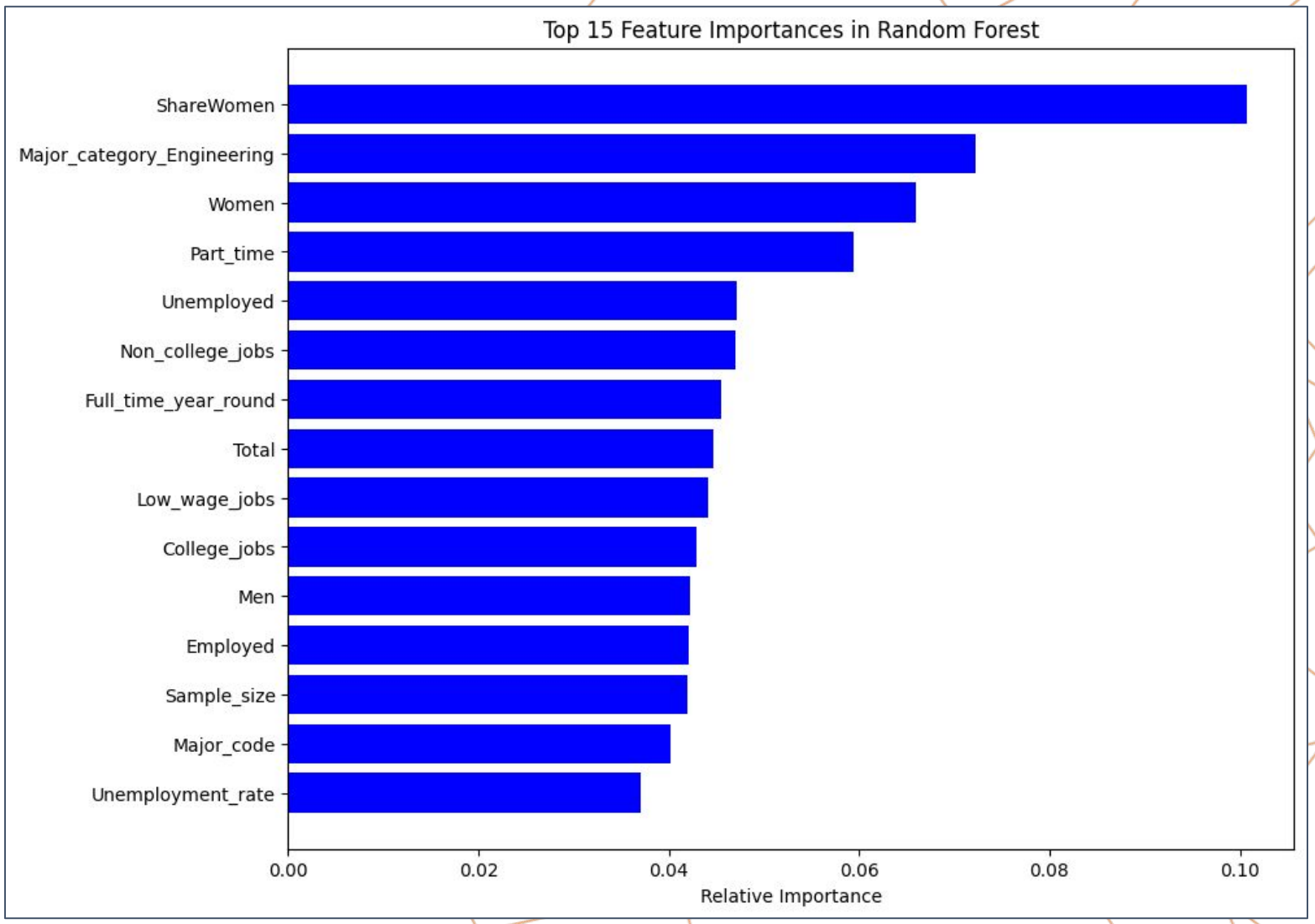
R^2 SCORE: 0.65



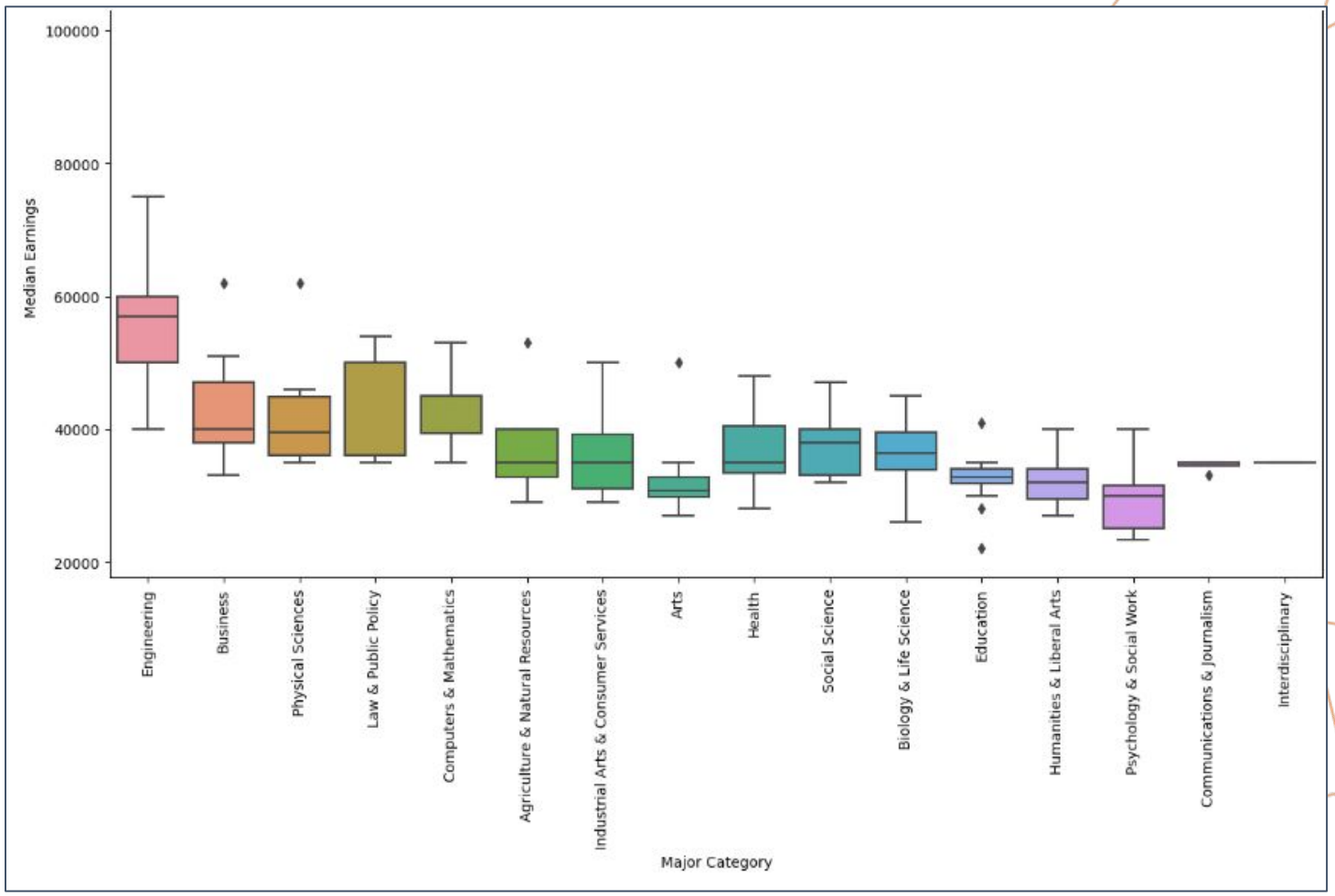
KEY FEATURES

Data is from 2010-2012

After part time the relative importance is relatively similar



Major Category vs Median Earnings

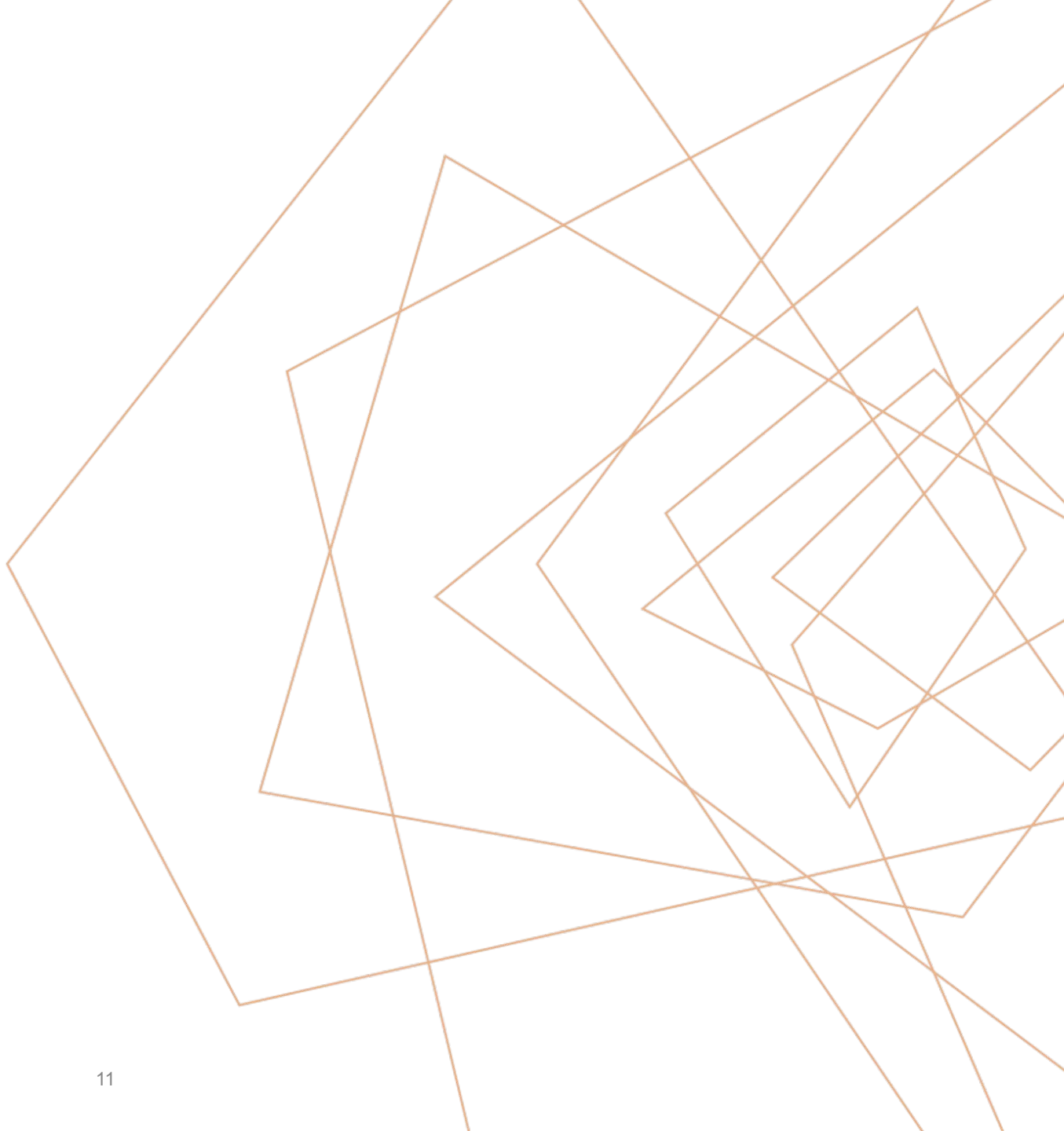


KEY TAKEAWAYS

Our model predicted that the Data Science major would earn $56,739 after being adjusted for inflation.

Predicted median earning in California is $89,363.

Average entry level Data Scientists salary is between $80k-$90k. via springboard.com



RESULTS AND CONCLUSIONS

Findings: The application of ML paradigms yielded findings that show the relationship between college major types, gender, and overall income.

Implications: The implications show the impact of the gender wage gap on the newly graduated job market.

Real-World Readiness: The model is only trained on 2011 data, could also include more features for future iterations.